

2º CICLO DE ESTUDOS

MESTRADO EM INFORMÁTICA MÉDICA

Construção de um modelo utilizando a Análise Discriminante para a deteção de Acidemia Fetal

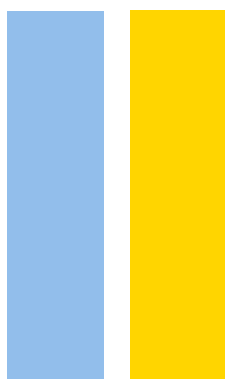
Tânia Raquel Martins Ferreira

M

2018

ORIENTADORA: Doutora Maria Luísa Morais Sousa Castro Guedes

COORIENTADORA: Doutora Andreia Sofia da Costa Teixeira



*“Ninguém escapa ao sonho de voar, de
ultrapassar os limites do espaço onde nasceu,
de ver novos lugares e novas gentes.
Mas saber ver em cada coisa, em cada pessoa,
aquele algo que a define como especial, um
objeto singular, um amigo - é fundamental.
Navegar é preciso, reconhecer o valor das
coisas e das pessoas, é mais preciso ainda!”*

Antoine de Saint-Exupéry

Agradecimentos

Primeiramente gostaria de agradecer à minha orientadora, a Doutora Maria Luísa Morais Sousa Castro Guedes e à minha co-orientadora, a Doutora Andreia Sofia da Costa Teixeira, por toda a paciência, pela disponibilidade sempre constante, pelos enormes conhecimentos transmitidos, orientações e comentários sempre construtivos e sábios, pela amizade e apoio em todos os momentos que formaram um pilar essencial para que este trabalho fosse possível.

Agradeço também ao Marcelo Santos, do departamento de Ciência de Computadores da Faculdade de Ciências da Universidade do Porto, por ter disponibilizado os cálculos dos índices não lineares dos traçados das frequências cardíacas fetais.

Um agradecimento aos meus amigos, que de uma forma direta ou indireta, contribuíram para a elaboração deste estudo. Um agradecimento especial à Catarina Varela pela amizade, força e constante incentivo neste momento importante da minha vida.

Por último, mas não menos importante, agradeço aos meus pais e irmão, pelo constante amor, suporte e incentivo ao longo de toda a minha vida académica. Muito obrigada pelo apoio incondicional.

Resumo

Esta dissertação tem como objetivo principal a criação de um modelo para a detecção antecipada de acidemia em fetos humanos.

Para tal, foram incluídos no estudo 146 bebés, nascidos entre 2005 e 2007. Foram analisados os traçados do batimento cardíaco fetal desses bebés, a partir da 37^a semana de gestação. Destes traçados, foram analisados índices lineares, do domínio do tempo, obtidos pela central de monitorização SisPorto®, e do domínio da frequência, calculados com o software MATLAB®. Foram também analisados índices não lineares, da área da compressão e da área da teoria da informação.

Aplicou-se a técnica multivariada de análise discriminante a um conjunto de treino, cujas variáveis são os índices lineares e não lineares de uma sub-amostra aleatória do conjunto dos traçados do batimento cardíaco fetal.

Esta dissertação permitiu perceber que dados dois conjuntos de variáveis compostas pela compressão, a entropia da amostra, a entropia aproximada e a análise espectral, no primeiro conjunto tanto pelo método de *variáveis independentes juntas* como pelo método *Stepwise*, foi possível classificar corretamente 85.3% dos casos. Já no segundo conjunto de variáveis o método de *variáveis independentes juntas* classificou corretamente 85.3% e através do método *Stepwise* classificou de forma correta 79.4%.

Em relação à classificação de novos casos, o método de *variáveis independentes juntas* obteve melhores resultados comparativamente ao método *Stepwise*.

Palavras-chave: Acidemia fetal; Índices lineares e não lineares; Análise discriminante.

Abstract

The aim of this dissertation was to create a model to an early detection of fetal acidosis.

One hundred and forty six FHR tracings, of newborns (whose delivery was between 2005 and 2007) with 37 or more weeks of gestation, were analyzed in this study.

Based on these tracings we analyzed time domain linear indices, that were acquired with SisPorto® system, and frequency domain, calculated with MATLAB® software. Non linear indices, from the area of compression and information teory, were also analysed.

Discriminate analysis multivariated technique was applied to a training set, obtained as a random sub-sample of the original FHR dataset. The independent variables in the analysis were the computed linear and non linear indices.

This dissertation allows us to know that taking to account two groups of variables, compound by compreesion, SampEn, ApEn and the spectral analysis, in the first one was possible to correctly classified 85,3% of the cases with both independent variables together and Stepwise methods. In the second group the independent variables together method classified 85,3% and Stepwise 79,4%.

Keywords: Fetal acidemia; Linear and nonlinear indices; Discriminant analysis.

Índice

Agradecimentos.....	5
Resumo.....	7
<i>Abstract</i>	9
Índice	11
Índice de tabelas.....	13
Índice de acrónimos	17
Introdução.....	19
1. Revisão da literatura	21
1.1. Acidose.....	21
1.2. Cardiotocografia	22
1.3. Análise discriminante	23
1.3.1. Objetivos da Análise Discriminante	23
1.3.2. Pressupostos da Análise Discriminante	23
1.4. Índices	28
1.5.1 Índices lineares.....	29
1.5.2. Índices não lineares	30
2. Metodologia.....	33
2.1. Objetivo	33
2.1.1. Objetivos específicos	33
2.2. Caraterização da amostra.....	33
2.3. Índices lineares.....	35
2.4. Índices não lineares	36
2.3. Métodos estatísticos	39
2.3.1. Análise de dados	39
2.3.2. Pré-processamento	39
2.3. Classificação	40

3. Resultados.....	41
3.1. Análise discriminante aplicada ao conjunto de treino.....	43
3.1.1. Método das <i>variáveis independentes juntas</i> aplicado ao primeiro conjunto de variáveis	43
3.1.2. Método <i>Stepwise</i> aplicado ao primeiro conjunto de variáveis	49
3.1.3. Método das <i>variáveis independentes juntas</i> aplicado ao segundo conjunto de variáveis	53
3.1.4. Método <i>Stepwise</i> aplicado ao segundo conjunto de variáveis.....	58
3.2. Classificação do conjunto de teste	61
4. Discussão	63
5. Conclusão e trabalho futuro.....	65
Referências	67
Anexos	69

Índice de tabelas

Tabela 1 - Características biométricas dos 146 recém-nascidos que constituem a amostra total. Fonte: elaboração própria.....	35
Tabela 2 - Grupos de variáveis analisadas para o modelo. Fonte: elaboração própria.....	37
Tabela 3 - Características biométricas do conjunto de treino e do conjunto de teste. Fonte: elaboração própria.	41
Tabela 4 - Resultados do teste para igualdade das distribuições no conjunto de treino - Teste Mann-Whitney. Apenas as variáveis significativas são apresentadas. Fonte: elaboração própria.	42
Tabela 5 - Média e desvio-padrão das variáveis do primeiro modelo de AD no grupo dos não acadêmicos, no grupo dos acadêmicos e nos dois em conjunto. Fonte: elaboração própria.	43
Tabela 6 - Lambda de Wilks das variáveis no modelo de AD e teste de igualdade das médias. Fonte: elaboração própria.....	44
Tabela 7 - Matriz de covariância e de correlação. Fonte: elaboração própria.....	45
Tabela 8 - Matriz de covariâncias em cada grupo do conjunto de treino. Fonte: elaboração própria.	45
Tabela 9 - Valor próprio. Fonte: elaboração própria.....	46
Tabela 10 - Lambda de Wilks da função discriminante do modelo de AD e teste de igualdade das médias. Fonte: elaboração própria.....	46
Tabela 11 - Coeficientes padronizados da função discriminantes e matriz de estrutura. Fonte: elaboração própria.....	47
Tabela 12 - Coeficientes da função discriminante de Fisher. Fonte: elaboração própria.....	47
Tabela 13 - Coeficientes das funções score de classificação. Fonte: elaboração própria.	48
Tabela 14 - Resultados de classificação original e com validação cruzada. Fonte: elaboração própria.	49
Tabela 15 - Variáveis incluídas/excluídas ^{a,b,c,d} no modelo recorrendo ao Lambda de Wilks. Fonte: elaboração própria.....	49
Tabela 16 - Variáveis incluídas na análise pelo método <i>Stepwise</i> . Fonte: elaboração própria.	50
Tabela 17 - Variáveis excluídas da análise pelo método <i>Stepwise</i> . Fonte: elaboração própria.	50
Tabela 18 - Valor próprio. Fonte: elaboração própria.....	50

Tabela 19 - Lambda Wilks. Fonte: elaboração própria.	51
Tabela 20 - Coeficientes padronizados da função discriminantes e matriz de estrutura. Fonte: elaboração própria.	51
Tabela 21 - Coeficientes da função de Fisher. Fonte: elaboração própria.	51
Tabela 22 - Coeficientes das funções de classificação. Fonte: elaboração própria.	52
Tabela 23 - Resultados de classificação original e com validação cruzada. Fonte: elaboração própria.	52
Tabela 24 - Média e desvio-padrão das variáveis do segundo modelo de AD no grupo dos não académicos, no grupo dos académicos e nos dois em conjunto. Fonte: elaboração própria.	53
Tabela 25 - Lambda de Wilks das variáveis no modelo de AD e teste de igualdade das médias. Fonte: elaboração própria.	54
Tabela 26 - Matriz de covariância e de correlação. Fonte: elaboração própria.	54
Tabela 27 - Matriz de covariâncias em cada grupo do conjunto de treino. Fonte: elaboração própria.	55
Tabela 28 - Valor próprio. Fonte: elaboração própria.	55
Tabela 29 - Lambda de Wilks da função discriminante do modelo de AD e teste de igualdade das médias. Fonte: elaboração própria.	55
Tabela 30 - Coeficientes padronizados da função discriminantes e matriz de estrutura. Fonte: elaboração própria.	56
Tabela 31 - Coeficientes da função discriminante de Fisher. Fonte: elaboração própria.	56
Tabela 32 - Coeficientes das funções score de classificação. Fonte: elaboração própria.	57
Tabela 33 - Resultados de classificação original e com validação cruzada. Fonte: elaboração própria.	57
Tabela 34 - Variáveis incluídas/excluídas ^{a,b,c,d} no modelo recorrendo ao Lambda de Wilks. Fonte: elaboração própria.	58
Tabela 35 - Variáveis incluídas pelo método <i>Stepwise</i> . Fonte: elaboração própria.	58
Tabela 36 - Variáveis excluídas pelo método <i>Stepwise</i> . Fonte: elaboração própria.	59
Tabela 37 - Valor próprio. Fonte: elaboração própria.	59
Tabela 38 - Lambda Wilks. Fonte: elaboração própria.	59
Tabela 39 - Coeficientes padronizados da função discriminantes e matriz de estrutura. Fonte: elaboração própria.	60
Tabela 40 - Coeficientes da função de Fisher. Fonte: elaboração própria.	60
Tabela 41 - Coeficientes das funções de classificação. Fonte: elaboração própria.	60

Tabela 42 - Resultados de classificação original e com validação cruzada. Fonte: elaboração própria.	61
Tabela 43 - Variáveis para os elementos do conjunto de teste. Fonte: elaboração própria.	61
Tabela 44 - Score das funções discriminantes para cada caso do conjunto de teste pelos dois métodos de AD: <i>variáveis independentes juntas versus Stepwise</i> . Fonte: elaboração própria.	62
Tabela 45 - Correlações de Spearman entre as variáveis consideradas significativas após o teste de Mann-Whitney. Fonte: elaboração própria.....	70
Tabela 46 - Teste para igualdade de distribuições da base de dados total – Teste Mann-Whitney. Fonte: elaboração própria.....	73

Índice de acrónimos

AD – análise discriminante

ApEn – entropia aproximada

bmp – batimentos por minuto

CTG – cardiotocografia

FCF – frequência cardíaca fetal

FIGO – Federação Internacional de Ginecologia e Obstetrícia

HF – frequências altas

Hz – hertz

II – índice de intervalo

LF – frequências baixas

LTI – irregularidade a longo prazo

mEq/L – miliequivalentes por litro

mmHg – milímetro de mercúrio

SampEn – entropia amostral

pH – potencial de hidrogénio

sd – desvio-padrão

sdFCF – desvio-padrão do sinal de FCF

SPSS – Statistical Package for the Social Sciences

STV – variabilidade de curto prazo

VLF – frequências muito baixas

Introdução

As tecnologias de informação têm-se revelado de extrema utilidade em diversas áreas, nomeadamente na área da saúde, onde se têm destacado pela sua importância no apoio clínico. Desde o início do século XX que se afirma a necessidade, no contexto hospitalar, em registar e organizar os dados clínicos referentes a cada episódio. Os avanços nas tecnologias de informação têm contribuído para o desenvolvimento de soluções informáticas hospitalares. Existe uma atualização das aplicações atuais, com o propósito de aumentar a eficácia das mesmas, ao passo que outras aplicações são desenvolvidas com novos requisitos e especificações, ficando ao dispor das diversas cadeias hospitalares e cooperando numa melhor prestação de cuidados de saúde [1].

Os fetos humanos, no decorrer do período de gestação, encontram-se expostos a diversos eventos que podem condicionar as trocas gasosas placentares, o que pode levar a uma diminuição da quantidade de oxigénio no sangue e a um aumento do dióxido de carbono (*hipoxemia*). Este aumento do dióxido de carbono resulta numa diminuição do pH do sangue fetal (*acidemia respiratória*), o que em condições normais é possível repor após retomar as trocas gasosas. A hipoxemia prolongada provoca um decréscimo do oxigénio, ao nível dos tecidos, o que faz com que as células recorram ao metabolismo anaeróbio para a produção de energia (*hipóxia*), podendo levar à acumulação tecidular de ácido láctico, denominada de *acidose metabólica*.

A hipóxia e a acidose metabólica prolongadas podem originar lesões irreversíveis em órgãos importantes do organismo ou até mesmo levar à morte, por asfixia ao nascimento. Em 2003, o Colégio Americano de Obstetrícia e Ginecologia (ACOG) e a Academia Americana de Pediatria (AAP) definiram a asfixia ao nascimento como uma situação de acidemia, hipóxia e acidose metabólicas suficientemente graves para causar lesões no feto [2, 3].

A incidência de morbimortalidade advinda de hipoxia é variável em todo o mundo. Nos países desenvolvidos ronda os 4 por 1000 nados-vivos, enquanto que nos países em desenvolvimento ronda os 33 por 1000 nados-vivos [4]. O principal meio para o médico avaliar o estado do feto, durante o trabalho de parto, é a *cardiotocografia* (CTG), que avalia a pressão uterina materna e a frequência cardíaca fetal.

O primeiro capítulo desta dissertação contém a revisão da literatura, onde são descritos conceitos fundamentais para a realização deste trabalho. O segundo capítulo consiste na descrição da metodologia do trabalho. O terceiro capítulo descreve os resultados obtidos através da análise discriminante. O quarto capítulo consiste na discussão dos resultados obtidos. No último capítulo são apresentadas as elações retiradas do estudo as limitações e dificuldades sentidas na realização do trabalho e são propostas algumas sugestões de trabalho futuro.

1. Revisão da literatura

1.1. Acidose

A *acidose* no feto consiste em todos os processos que levam a uma alta concentração de iões de hidrogénio no sangue arterial do cordão umbilical, causando acidemia. *Acidemia* ocorre quando se verifica um nível anormalmente baixo do potencial de hidrogénio (pH). O *pH* é uma escala numérica adimensional utilizada para especificar a acidez ou basicidade de uma solução aquosa. O pH é definido pelo simétrico do logaritmo na base dez da concentração de iões de hidrogénio em mol/L (em moles por litro).

No feto, o intervalo normal do pH na veia umbilical varia de 7.25 a 7.45 (extremos inclusive), enquanto que na artéria umbilical varia de 7.20 a 7.38 (extremos inclusive). Diz-se que a *acidose fetal* acontece quando um vaso umbilical apresenta um pH inferior a 7.20 e um excesso de base inferior a -8. O excesso de base é definido como sendo a quantidade de ácido forte que deve ser adicionado a cada litro de sangue oxigenado para que este retorne a um pH de 7.4, à temperatura de 37°C e à pressão parcial de dióxido de carbono (pCO₂) de 40 mmHg [5]. Entende-se como excesso de base normal o intervalo entre -2 e +2 mEq/L (miliequivalentes por litro).

O feto depende da placenta para efetuar as trocas de oxigénio e de dióxido de carbono, e perturbações em qualquer um destes gases podem causar sofrimento fetal, o que, apesar dos mecanismos compensatórios, pode levar à acidemia [6]. As causas da hipóxia fetal (privação de oxigénio no feto) podem ser classificadas como *hipóxia fetal aguda* (se tiver uma duração de horas) ou *crónica* (se tiver uma duração de dias) e podem ser divididas de acordo com a origem, isto é, maternas, placentárias ou fetais e as suas consequências dependem da gravidade, da duração e também das condições do feto antes do parto [6].

1.2. Cardiotocografia

A *cardiotocografia* (CTG) é um importante meio de diagnóstico na gravidez. Consiste no registo dos batimentos cardíacos fetais (FCF) em batimentos por minuto (bpm), das contrações uterinas em mmHg e dos movimentos fetais (presentes/ausentes).

Os sinais da CTG são difíceis de analisar, pois as limitações inerentes à visão humana dificultam a análise visual do mesmo. É de referir também, a dificuldade em obter um consenso entre os profissionais de saúde na análise e leitura, tanto dos traçados como dos parâmetros da CTG, como, por exemplo, a determinação da linha de base, a classificação de acelerações e de desacelerações [7].

Estudos realizados em 1997, no Reino Unido, revelaram que 50% das mortes oriundas de hipóxia *intraparto* (detetada durante o trabalho de parto) poderiam ter sido evitadas, uma vez que foram fruto de uma falta de reconhecimento de FCF anormais, falhas de comunicação e de interpretação entre profissionais de saúde ou, atrasos em tomadas de decisão imediatas [8]. A análise dos traçados de FCF é realizada a nível mundial, maioritariamente, por profissionais de saúde, o que suscita uma ampla variação de observações, tanto no que diz respeito à interpretação [9] como à identificação dos traçados individuais [10].

O desenvolvimento de sistemas de análise automáticos tem como objetivo tentar diminuir ou eliminar algumas das dificuldades e ambiguidades da análise visual dos traçados da CTG [11]. Com o objetivo de superar as limitações advindas da análise visual das CTG [12], surgiu em 1987, o desenvolvimento do primeiro sistema original de aquisição, processamento e análise automática da CTG [13].

A análise computadorizada dos traçados de FCF tem continuado a ser desenvolvida ao longo das últimas décadas, com a finalidade de melhorar a fraca reprodutibilidade da análise visual. A maioria dos sistemas está focado em analisar os traçados *anteparto* (antes do início do trabalho de parto) [14] denotando-se grandes limitações quando se trata de traçados *intraparto* [15]. Os traçados *intraparto* possuem algumas dificuldades para a análise computadorizada, normalmente relacionadas com o tamanho dos mesmos, como as perdas de sinal, sendo necessária uma maior memória para o armazenamento.

1.3. Análise discriminante

A análise discriminante (AD) é uma técnica multivariada utilizada para analisar a relação entre uma variável dependente não métrica e variáveis independentes métricas (variáveis explicativas ou também denominadas de preditoras). O objetivo é que através das variáveis independentes, seja possível distinguir os grupos (categorias) da variável dependente, ou seja, esta pretende descobrir as características (variáveis explicativas) que distinguem os membros de um grupo dos de outro, de forma a que conhecidas as características de um novo caso (indivíduo), seja possível classificar a que grupo da variável dependente pertencerá [16].

Esta ferramenta foi inicialmente desenvolvida na área da botânica, de forma a distinguir grupos de plantas com base no tamanho e tipos de folhas, o que permitia, posteriormente, classificar novas espécies de plantas [17]. Rapidamente, esta técnica passou a ser utilizada em outras áreas como as ciências sociais e humanas, a economia, as biomédicas, entre outras [18].

Quando a variável dependente possui 2 grupos, utiliza-se a Análise Discriminante Simples (ADS); quando a variável dependente possui 3 ou mais grupos, utiliza-se a Análise Discriminante Múltipla (ADM). Nesta dissertação, apenas será abordada a Análise Discriminante Simples, doravante abreviada por AD.

1.3.1. Objetivos da Análise Discriminante

Os objetivos da AD são os seguintes:

- Identificar as variáveis explicativas que melhor discriminam entre dois ou mais grupos de indivíduos;
- Criar funções discriminantes que representem as diferenças entre os grupos;
- Utilizar essas funções discriminantes para classificar novos indivíduos nos grupos.

1.3.2. Pressupostos da Análise Discriminante

A utilização desta técnica gera taxas de erros mínimos na classificação quando são verificados os seus pressupostos, que, de acordo com [19], são:

- **Normalidade multivariada das variáveis explicativas:** as variáveis independentes devem possuir normalidade multivariada ou, pelo menos, a maioria deve possuir normalidade univariada.
- **Homogeneidade das variâncias dos grupos:** os grupos devem possuir variâncias semelhantes. Se um dos grupos possui uma maior variância relativamente aos outros, os novos casos tenderão a ser classificados nesse grupo.
- **Inexistência de outliers:** a técnica é comprometida com a existência de valores discrepantes para cada variável independente.
- **Relação entre casos e grupos:** deve existir pelo menos 20 casos para cada grupo. O número de casos em cada grupo deve ser aproximadamente igual. Quando se utilizam k variáveis independentes, o tamanho mínimo da amostra deve ser $5k$.
- **Multicolinearidade:** a técnica funciona melhor se as variáveis independentes possuem correlações baixas. Correlações altas indicam que as variáveis independentes são redundantes e uma delas deve ser excluída.

Em relação à normalidade multivariada das variáveis, a combinação linear das variáveis explicativas deve apresentar uma distribuição normal. Caso ocorra uma violação desse pressuposto, a AD poderá causar distorções no modo como o investigador faz a avaliação, principalmente se a amostra de cada grupo for pequena. Contudo, se esta violação ocorrer somente pela existência de assimetria da distribuição, a aplicação da técnica não sofrerá alteração. É importante perceber que, se a distribuição não for mesocúrtica, isto é, não apresentar uma curvatura semelhante à da distribuição normal, a aplicação da AD será prejudicada, sendo pior no caso em que a distribuição é platicúrtica, isto é, a curva apresenta-se mais “achatada” que a da distribuição normal [20].

O pressuposto da homogeneidade das variâncias dos grupos é verificado através da estatística Box's M (M de Box), que pode ser sensível ao tamanho da amostra e ao incumprimento da hipótese de distribuição normal multivariada [21]. Hair *et al.* destacam que é essencial definir o tamanho correto da amostra a ser estudada, já que esta técnica é muito sensível no que diz respeito ao tamanho da amostra em relação ao número de variáveis explicativas. Portanto, não deve haver uma grande diferença nas dimensões entre os grupos, como já foi dito anteriormente [16].

A AD gera m funções discriminantes (Z_m), que são combinações lineares das variáveis explicativas, que ampliam a discriminação dos grupos da variável dependente.

Estas funções são também conhecidas por funções discriminantes de Fisher. Nem toda a função discriminante é significativa. Normalmente, apenas m funções são significativas. O número máximo de funções discriminantes é dado por $\min(g-1; k)$, onde g é o número de grupos da variável dependente e k é o número total de variáveis independentes. A função discriminante é definida por:

$$Z_m = B_{m0} + B_{m1}X_1 + B_{m2}X_2 + \dots + B_{mk}X_k$$

Sendo B_{m0} uma constante, X_1, X_2, \dots, X_k são as variáveis independentes e $B_{m1}, B_{m2}, \dots, B_{mk}$ são os m-ésimos coeficientes da discriminante das variáveis independentes, que maximizam a distância entre as médias dos grupos e minimizam a variância dentro destes. Seja $T = B+W$, em que B é a matriz da soma dos quadrados entre os grupos das variáveis independentes e produto cruzado da matriz X com k variáveis e W é a matriz da soma dos quadrados dentro dos grupos das variáveis independentes e produto cruzado da matriz X com k variáveis. Considere-se a equação:

$$W^{-1}BV = \lambda V \rightarrow (W^{-1}B - \lambda I)V = 0$$

Em que I é a matriz identidade; V é a matriz de valores discriminantes intermediários e λ é o valor próprio. Cada valor próprio é encontrado resolvendo o sistema:

$$|W^{-1}B - \lambda I| = 0$$

Os valores discriminantes intermediários da matriz V possuem o formato $V_m = v_{m0} + v_{m1}X_1 + \dots + v_{mk}X_k$, onde:

$$v_{m0} = -\sum_k v_{mk}\bar{X}_k$$

A matriz V possui, para cada valor próprio, infinitas soluções. No entanto, os coeficientes discriminantes intermediários que originam os coeficientes da função discriminante são encontrados quando a razão entre a soma dos quadrados entre grupos e a soma dos quadrados dentro dos grupos é máxima.

Os coeficientes discriminantes intermediários são calculados como:

$B_{mk} = v_{mk}\sqrt{n-g}$ e $B_{m0} = -\sum_k B_{mk}\bar{X}_k$, sendo n o tamanho da amostra; g o número de grupos e \bar{X}_k a média dos valores da variável independente k.

O centróide (coordenada central) do grupo g é obtido quando se utiliza a média de cada variável independente k que pertença ao grupo g em cada função discriminante:

$$Z_{gm} = B_{m0} + B_{m1}\bar{X}_1 + \dots + B_{mk}\bar{X}_k$$

ou

$$Z_{gm} = \bar{Z}_m, \forall Z_m \in g$$

Onde Z_{gm} é a m -ésima coordenada do centróide g ; \bar{Z}_m é a média dos valores da função discriminante para os casos de X_k que pertençam a g e \bar{X}_k é a média da variável independente k .

A partir de testes estatísticos, como o Lambda de Wilks, a correlação canónica, o teste Qui-quadrado, o teste M de Box e os valores próprios, a AD faz o conhecimento das variáveis que mais se destacam na discriminação dos grupos. O Lambda de Wilks para variáveis independentes (também denominado de Estatística-U) varia entre 0 e 1 e, para cada variável independente, verifica a existência da diferença de médias entre grupos. Quanto mais próximo de 0, mais discriminante é a variável independente. O seu valor pode ser transformado numa distribuição F, sendo possível testar a hipótese nula de que a média dos grupos é igual. Os valores próprios indicam, para cada função discriminante, o percentual total de variância explicada para cada função, ou seja, o quão importante é a função para que se consiga discriminar os grupos [22]. A correlação canónica mede a correlação entre cada função discriminante e os grupos. Quanto mais próximo de 1 é o valor da correlação canónica, mais discriminante é a função. O quadrado da correlação canónica indica o total de variância explicada pela função discriminante. O teste Qui-quadrado de Lambda de Wilks testa globalmente as funções discriminantes, e a cada iteração, uma função é retirada, até que reste apenas a última. A hipótese nula deste teste averigua se as funções são significativas para a discriminação dos grupos [20]. O teste M de Box verifica se a matriz de variância-covariância é igual entre os grupos. Este teste não é considerado robusto e pode ser substituído pela tabela de determinante de log, que indica a homogeneidade se os valores do determinante para cada grupo forem similares. Os coeficientes discriminantes padronizados permitem verificar a importância relativa de cada variável independente. A matriz de estrutura auxilia na interpretação da contribuição que cada variável fornece para cada função discriminante, uma vez que apresenta as correlações entre as variáveis explicativas e as funções discriminantes canónicas padronizadas.

No SPSS® (*Statistical Package for the Social Sciences*) existem dois métodos para a elaboração das funções discriminantes:

- O método das *variáveis independentes juntas*: todas as variáveis explicativas são incluídas, conjuntamente no modelo, mesmo quando uma ou mais não são significativas.
- O método *Stepwise*: apenas as variáveis significativas são incluídas, passo a passo.

Esta técnica começa por selecionar as variáveis que mais contribuem para distinguir os grupos e em seguida vai incluindo ou excluindo as variáveis discriminantes na função discriminante.

Após a escolha do modelo que será utilizado, o SPSS® apresenta um conjunto de opções, permitindo ao investigador escolher aquelas que mais se enquadram na sua análise. A primeira opção denominada estatística, permite selecionar qual o coeficiente da função a apresentar. É possível escolher entre o de Fisher, que apresenta os coeficientes das funções score discriminantes de cada grupo ou o *não padronizado*, que expressa os coeficientes das funções discriminantes de Fisher com a origem localizada na grande média (centróide dos centróides). Outra opção é a classificação, que permite escolher se a probabilidade de um *indivíduo* ser colocado em dado grupo é igual em todos os grupos ou, se a probabilidade é calculada a partir das dimensões dos grupos. O método *Stepwise*, para além destas opções, fornece ainda um conjunto de opções específicas para a seleção das variáveis explicativas a integrar o modelo:

- Lambda de Wilks: esta opção seleciona as variáveis que minimizam, significativamente, o valor do Lambda de Wilks do modelo.
- Variância inexplicada: esta opção seleciona as variáveis que minimizam, significativamente, a variância não explicada entre os grupos.
- Distância de Mahalanobis: é utilizada para garantir a separação de todos os grupos. Seleciona as variáveis baseada nos casos de cada uma. Variáveis com muitos casos com valores discrepantes são excluídas.
- Menor razão F: consiste na transformação da distância de Mahalanobis numa razão F entre os grupos. Esta transformação considera as diferentes dimensões entre os grupos, fazendo com que grupos maiores tenham maior peso na análise. Esta opção seleciona as variáveis que maximizam a menor razão F para a distância de Mahalanobis.

- V de Rao: também conhecido como Traço de Lawley-Hotellingue que mede a distância entre grupos. Uma variável é selecionada se aumentar significativamente o valor de V de Rao. Não procura maximizar a homogeneidade dentro dos grupos, mas sim a heterogeneidade entre eles [18].

Uma das principais perguntas que se coloca é se devemos utilizar a AD mesmo quando os pressupostos não se verificam. A principal preocupação está na utilização dos resultados e na tolerância dos erros. O problema pode ser ultrapassado se a decisão tiver como base os resultados do Lambda de Wilks e as correlações canônicas. De certo modo, quando a percentagem de classificações corretas é elevada, podem servir sempre como um bom modelo de descrição da situação real, não sendo grave a violação dos pressupostos. Já no caso dessa percentagem ser baixa, não é possível prever se tal se deve ao fato de não terem sido respeitados os pressupostos ou se estamos diante de um conjunto de variáveis com fraco poder para discriminar os grupos.

Resumidamente, a análise discriminante é um método estatístico exaustivo para classificar indivíduos ou objetos em grupos mutuamente exclusivos, com base num conjunto de variáveis independentes. São criadas combinações lineares dessas variáveis que discriminam entre grupos definidos *a priori*, de modo a que seja minimizada a probabilidade de erro de uma incorreta classificação *a posteriori* [16].

1.4. Índices

Os dados utilizados neste estudo são sinais, traçados de batimento cardíaco fetal, ou seja, são séries temporais. Em geral, para se conseguir aplicar classificação a este tipo de dados, é necessário extrair índices desses sinais, que serão as variáveis em estudo. Existem dois tipos de índices usualmente utilizados em estudos que envolvem séries temporais: índices lineares e índices não lineares. Exemplos de índices lineares são os baseados em cálculos estatísticos ou obtidos pela análise espectral. Exemplos de índices não lineares são os baseados na Teoria do Caos, na Teoria da Informação ou em complexidade.

Gonçalves *et al.* realizaram, em 2006, dois estudos. O objetivo do primeiro era avaliar o efeito do modo de aquisição do sinal de FCF (interno e externo) nos índices lineares e não lineares [22]. No segundo estudo calcularam índices lineares e não lineares das frequências cardíacas e avaliaram as mudanças destes índices em diferentes janelas de

tempo (minutos) antes do parto, com o objetivo de determinarem a capacidade destes índices em distinguir os fetos normais dos fetos com acidemia [23].

1.5.1 Índices lineares

Os métodos convencionais para a extração de características para a análise de sinais biológicos utilizam normalmente métodos baseados no tempo e na frequência, uma vez que a teoria por detrás desses métodos assume que estes sinais são lineares [24].

No domínio da frequência, os índices lineares, permitem a decomposição do FCF em diferentes segmentos. Em ambos os trabalhos referidos anteriormente, Gonçalves *et al.*, utilizam os índices no domínio das frequências, para a obtenção das potências das várias gamas de frequência. Quatro categorias foram definidas com base em gamas de frequência [25]: *frequência muito baixa*, VLF: (0 - 0.03 Hz); *frequência baixa*, LF: (0.03 - 0.15 Hz); *frequência de movimento*, MF: (0.15 - 0.50 Hz) e *frequência alta*, HF: (0.50 - 1.00 Hz). Foram também considerados a TP (potência total obtida pela soma das frequências das várias bandas, toda a área do espectro) e o rácio $LF/(MF + HF)$. Um outro estudo realizado por Chung *et al.* utilizou uma gama de frequências diferente: VLF: (0 - 0.04 Hz), LF: (0.04 - 0.15 Hz), HF: (0.15 - 0.4 Hz) e o rácio LF/HF [26]. Os limites das gamas de frequência variam ligeiramente na literatura existente, mas os índices lineares no domínio da frequência têm obtido resultados promissores para a identificação prévia da acidemia [26]. Relativamente ao domínio do tempo, Gonçalves *et al.*, consideraram os seguintes índices lineares [23]: média do sinal de FCF (mFCF), desvio-padrão de FCF (sdFCF), irregularidade a longo prazo (LTI), variabilidade de curto prazo (STV) e índice de intervalo (II), que avalia a variabilidade a curto prazo da FCF levando em conta a variabilidade a longo prazo. Diz-se que a variabilidade curta (STV) é anormal quando a diferença entre frequências cardíacas adjacentes é inferior a 1 bpm. Diz-se que a variabilidade longa (LTV) é anormal quando a diferença entre o valor máximo e mínimo das frequências cardíacas não excede os 5 bpm, considerando uma janela temporal de 60 segundos.

Também têm sido feitos estudos, no que diz respeito à análise dos traçados completos dos batimentos cardíacos fetais pelo sistema Omniview-SisPorto®, como Ayres-de-Campos *et al.* referem em [27, 28]. Este *software* é a mais recente versão [29] de um sistema criado para analisar sinais anteparto e intraparto, que produz alertas quando os índices lineares calculados a partir do traçado de FCF apresentam valores suspeitos [29].

O Omniview-SisPorto® calcula a linha de base das frequências cardíacas usando um algoritmo complexo desenvolvido para identificar o significado das frequências cardíacas durante a gravidez. O sistema devolve o número de acelerações, sendo definidas pelo aumento da frequência cardíaca acima da linha de base, durante 15 a 120 segundos e atingindo o “pico” aos 15 bpm; o número de desacelerações, definidas pelo decréscimo da frequência cardíaca abaixo da linha de base, durante pelo menos 15 segundos e com uma amplitude acima dos 15 bpm. Desacelerações leves nos últimos 60 minutos são classificadas, seguindo as guidelines FIGO (Federação Internacional de Ginecologia e Obstetrícia), como *suaves* se tiverem uma duração inferior a 120 segundos, *prolongadas* se durarem entre 120 a 300 segundos e *severas* se excederem 300 segundos. Também produz a percentagem da perda de sinal e as contrações uterinas; a linha basal da FCF; as contrações nos últimos 60 minutos e o índice saltatório nos últimos 60 minutos. Ainda segundo as guidelines FIGO, o software reporta repetidas desacelerações nos últimos 20 minutos, quando ocorrem mais de 3 em 10 minutos e acompanhadas de 80% de contrações ou quando a sua duração excede 50% do traçado [27].

A escala ou índice de Apgar consiste na avaliação de 5 sinais objetivos (cor da pele, pulsação arterial, irritabilidade reflexa, atividade e esforço respiratório) do recém-nascido, atribuindo-se a cada um dos sinais uma pontuação de 0 a 2. Esta avaliação é efetuada geralmente duas vezes, no primeiro e no quinto minuto após o nascimento e é utilizado para avaliar a adaptação do recém-nascido à vida extrauterina. A quantificação das acelerações e das STV e LTV anormais, recorrendo ao Omniview-SisPorto® permite uma boa previsão deste índice [27].

1.5.2. Índices não lineares

Pincus *et al.* adaptou a *entropia de Kolmogorov-Sinai*, cujo objetivo é quantificar a regularidade e aleatoriedade de um sistema e desenvolveu a *entropia aproximada* (ApEn). A ApEn tem como objetivo efetuar uma quantificação da irregularidade de séries temporais, por meio da similaridade dos vetores próximos uns dos outros [30]. Este método reflete a probabilidade condicional de que dadas duas sequências de “*m*” pontos de dados consecutivos, que são semelhantes entre si, permaneçam semelhantes quando mais um ponto consecutivo é incluído. Sendo que semelhante significa que o valor de uma ação específica da distância é inferior a “*r*”. Assim, a noção de entropia depende dos parâmetros N, m e r que devem ser fixos para cada cálculo, onde N é o comprimento da série

temporal, m é o comprimento das sequências a serem comparadas e r é o limite de tolerância *versus* o desvio-padrão da série de dados a ser calculada [31]. A entropia aproximada apresenta baixa sensibilidade para ruídos e consegue demonstrar um comportamento determinístico e estocástico. Porém, existe uma grande relação entre o tamanho dos dados e a incongruência nos resultados [32]. O conceito de *entropia amostral* (SampEn) surgiu mais tarde, no ano de 2000, por Richman *et al.*, com o objetivo de reduzir o viés da ApEn, ao considerar que um vetor seja próximo a ele mesmo, isto é, não contabiliza as auto-correspondências (*self-matches*), permitindo que algumas limitações existentes na entropia aproximada fossem solucionadas [31, 33]. A SampEn foi posteriormente empregue na análise de sinais biomédicos como o Eletroencefalograma (EEG) [34].

A entropia aproximada e a entropia amostral são medidas não lineares obtidas através da estimativa direta do sinal, capazes de quantificar a complexidade do mesmo (ou irregularidade) de forma robusta em curtos segmentos [35]. Para cada uma destas medidas, são estabelecidos valores para m e r . Gonçalves *et al.* num estudo em 2006, consideraram para o parâmetro r os valores $0.10*sd$, $0.15*sd$ e $0.20*sd$, onde sd representa o desvio-padrão do sinal, e o valor 2 foi considerado para o parâmetro m . Nesse estudo mostraram que a progressão do trabalho de parto nos últimos minutos, tanto nos fetos normais como nos fetos com acidemia, está associada com um aumento significativo dos índices lineares do domínio da frequência, e com uma diminuição significativa dos índices não lineares. Utilizando uma combinação dos índices lineares e não lineares, foi possível separar os fetos com acidemia moderada dos fetos com acidemia severa. Estes resultados evidenciam o interesse clínico nos índices lineares e não lineares.[23]. Estes valores dos parâmetros r e m são usualmente escolhidos pois produzem uma boa validade estatística da entropia aproximada, fundamentada por cálculos que incluem análises teóricas e aplicações clínicas [36].

Outro exemplo de índices não lineares são os compressores. Em 2006, Costa-Santos *et al.* efetuaram um estudo baseado em 31 traçados de frequência cardíaca fetal (FCF), adquiridos antes do parto em 4 hospitais. Neste trabalho, os autores utilizaram uma medida de compressão (*Normalized Compression Distance, NCD*) como distância para o agrupamento hierárquico dos traçados, em classes, utilizando compressores fixos como gzip, bzip2 e PPMZ. Esta medida de compressão não requer conhecimento do histórico clínico e pretende encontrar padrões nos dados, sem necessidade de definir parâmetros. O objetivo

do trabalho era identificar e agrupar os traçados de FCF anormais, com diferentes padrões. No primeiro nível de agregação, o método agregou os traçados com a mesma taxa de amostragem. No segundo nível de agregação, o objetivo foi alcançado já que os traçados patológicos ou suspeitos foram corretamente agregados numa mesma classe, sendo assim separados dos traçados sem patologias [37].

2. Metodologia

2.1. Objetivo

Esta dissertação tem como objetivo elaborar, a partir de um conjunto de dados reais de frequência cardíaca fetal (FCF), um modelo, recorrendo à análise discriminante, que permita a deteção antecipada de fetos com acidemia. Serão analisados três grupos diferentes de variáveis, obtidas dos sinais de batimento cardíaco fetal, para a possível inclusão no modelo de previsão: um grupo de índices não lineares, nomeadamente, medidas de entropia e de compressão; um grupo constituído por índices lineares no domínio do tempo, calculados pelo sistema Omniview-SisPorto® [27]; e, um terceiro grupo de índices no domínio da frequência do sinal [26].

2.1.1. Objetivos específicos

Os objetivos específicos desta dissertação são os seguintes:

- Descrição detalhada das variáveis que constituem a base de dados e caracterização da amostra;
- Familiarização com os índices lineares no domínio do tempo e no domínio da frequência, e com os índices não lineares do FCF;
- Determinação da função discriminante que melhor classifica novos dados, tendo como variáveis os índices lineares e não lineares .

2.2. Caracterização da amostra

A amostra é constituída pelo mesmo conjunto de dados que Costa *et al.* utilizaram em “*Prediction of neonatal acidemia by computer analysis of fetal heart rate and ST event signals*” [38]. Nesta base de dados foram incluídas gravidezes normais com mais de 36 semanas de

gestação, sem malformações fetais, com fase ativa de parto e indicação para monitorização fetal. Foram excluídos os casos em que os traçados tinham menos de 60 minutos, os casos nos quais ocorreu uma perda de sinal superior a 15% na última hora que antecede o parto, os que apresentam complicações na oxigenação fetal, complicações no caso do uso de anestesia e valores no sangue do cordão umbilical considerados inadequados. Quando, por razões práticas relacionadas com o tempo necessário para aplicação de ventosa ou preparação para cesariana, o intervalo entre o fim do traçado e o parto vaginal excede os 5 minutos ou, no caso da cesariana exceder os 20 minutos, os pacientes em causa foram também excluídos. De um total de 193 casos foram excluídos 45: 19 por desconexão do eletrodo do couro cabeludo, o que condicionou num intervalo longo do traçado ou perda de sinal; 21 com valores inadequados dos gases do cordão umbilical e 5 pela insuficiente duração do traçado [38]. Dois casos não foram considerados, uma vez que o traçado de batimento cardíaco fetal não tinha sido disponibilizado à data do início deste trabalho.

Para a realização do presente trabalho de investigação foram incluídos 146 traçados de recém-nascidos, com uma taxa de amostragem de 4Hz. Adicionalmente outras características biométricas relativas ao recém-nascido foram disponibilizadas e encontram-se sumariadas na seguinte tabela.

Tabela 1 - Características biométricas dos 146 recém-nascidos que constituem a amostra total. Fonte: elaboração própria.

Caraterísticas		Amostra total
Sexo, n(%)	Feminino	74 (50.7)
	Masculino	72 (49.3)
Tipo de parto, n(%)	Normal	72 (49.3)
	Cesariana	31 (21.2)
	Instrumental	43 (29.5)
pH artéria umbilical, n(%)	≤ 7.15	20 (13.7)
	> 7.15	126 (86.3)
Idade gestacional, n(%)	36	2 (1.4)
	37	10 (6.8)
	38	37 (25.3)
	39	44 (30.1)
	40	39 (26.7)
	41	14 (9.6)
Índice de Apgar, $\bar{x} \pm sd$	1º min	8.40 \pm 1.02
	5º min	9.58 \pm 0.75
	10º min	9.88 \pm 0.41
Peso (gramas), $\bar{x} \pm sd$		3232.55 \pm 389.67
Duração dos traçados (min), $\bar{x} \pm sd$	Mínimo: 60	210.77 \pm 129.63
	Máximo: 780	

Como é possível verificar na Tabela 1, existem cerca de seis vezes mais casos com pH da artéria do cordão umbilical acima de 7.15 (casos normais) do que abaixo desse valor de corte (casos acidêmicos).

2.3. Índices lineares

No que diz respeito aos índices lineares, serão analisados índices no domínio do tempo, calculados pelo sistema Omniview-SisPorto® [27], e índices no domínio da frequência do sinal [26]. Dos índices obtidos através do sistema Omniview-SisPorto® [27], calculados a partir do sinal completo, os índices incluídos no estudo foram os seguintes: linha de base e basal da FCF, acelerações nos últimos 60 minutos, contrações nos últimos 60 minutos, curta variabilidade anormal nos últimos 60 minutos, média da curta variabilidade nos últimos 60 minutos, longa variabilidade anormal nos últimos 60 minutos, média da longa variabilidade nos últimos 60 minutos, índice saltatório nos últimos 60 minutos e desacelerações moderadas nos últimos 60 minutos.

A análise espectral permite capturar as frequências de um sinal e desempenha um papel fundamental no estudo das características de uma série, uma vez que permite a detecção de periodicidades escondidas nos dados. Para o cálculo dos índices no domínio da frequência, os sinais de batimento cardíaco fetais foram normalizados e a densidade espectral foi obtida pelo método do periodograma através do software MATLAB®. Segundo Brillinger (1981), o periodograma é um estimador espectral assintoticamente consistente, não enviesado e não correlacionado [39]. A potência foi então obtida para as seguintes gamas de frequência: VLF (0 – 0.04 Hz); LF (0.04 – 0.15 Hz); HF (0.15 – 0.4 Hz) e o rácio LF/HF foi calculado [26]. Uma lista de todas as variáveis analisadas para inclusão no modelo de análise discriminante encontra-se apresentada na Tabela 2.

2.4. Índices não lineares

Relativamente aos índices não lineares, serão analisadas a entropia e a compressão, que foram calculadas sobre subsegmentos dos últimos 50 minutos do traçado original, os primeiros 5, 10 e 20 minutos dos últimos 50 minutos e os últimos 5, 10 e 20 minutos desses últimos 50 minutos (Tabela 2).

Relativamente às variáveis baseadas na noção de entropia, foram aplicadas a entropia aproximada e a entropia amostral. Em ambas foi considerado o valor 2 para a dimensão embutida (m) e, no caso do parâmetro r foram considerados os valores $0.10 \cdot sd$, $0.15 \cdot sd$ e $0.20 \cdot sd$ [23]. No caso da compressão, foi utilizada a taxa de compressão, que consiste na razão entre o tamanho do sinal comprimido e o original, multiplicada por 100. Foram utilizados os compressores mais comuns: Brotli¹, Bzip2², Gzip³, Lzma⁴, Paq8l⁵, Ppmd⁶ (Tabela 2).

¹ Brotli foi desenvolvido entre 2013 e 2016 pelos funcionários do Google.

² Bzip2 foi desenvolvido por Julian Seward 1996. (www.bzip.org)

³ Gzip foi desenvolvido por Jean-loup Gailly e Mark Adler em 1992. (www.gzip.org)

⁴ Lzma foi desenvolvido entre 1996 e 1998 por Igor Pavlov.

⁵ Paq8l foi lançado em 2006 por Pavel L. Holoborodko.

⁶ Ppmd foi desenvolvido por J.Cleary e I.Witten em 1984.

Tabela 2 - Grupos de variáveis analisadas para o modelo. Fonte: elaboração própria.

Grupo	Abreviatura da variável	Descrição da variável
Índices não lineares	Entropia Aproximada	ApEn 0.10 50min
		r = 0.10*sd, dos últimos 50 minutos
		ApEn 0.15 50min
		r = 0.15*sd, dos últimos 50 minutos
		ApEn 0.20 50min
		r = 0.20*sd, dos últimos 50 minutos
		ApEn 0.10 p5min
		r = 0.10*sd, dos primeiros 5 minutos
		ApEn 0.10 p10min
		r = 0.10*sd, dos primeiros 10 minutos
		ApEn 0.10 p20min
		r = 0.10*sd, dos primeiros 20 minutos
		ApEn 0.15 p5min
		r = 0.15*sd, dos primeiros 5 minutos
		ApEn 0.15 p10min
		r = 0.15*sd, dos primeiros 10 minutos
		ApEn 0.15 p20min
		r = 0.15*sd, dos primeiros 20 minutos
		ApEn 0.20 p5min
		r = 0.20*sd, dos primeiros 5 minutos
		ApEn 0.20 p10min
		r = 0.20*sd, dos primeiros 10 minutos
		ApEn 0.20 p20min
		r = 0.20*sd, dos primeiros 20 minutos
		ApEn 0.10 u5min
		r = 0.10*sd, dos últimos 5 minutos
		ApEn 0.10 u10min
		r = 0.10*sd, dos últimos 10 minutos
		ApEn 0.10 u20min
		r = 0.10*sd, dos últimos 20 minutos
		ApEn 0.15 u5min
		r = 0.15*sd, dos últimos 5 minutos
		ApEn 0.15 u10min
		r = 0.15*sd, dos últimos 10 minutos
		ApEn 0.15 u20min
		r = 0.15*sd, dos últimos 20 minutos
		ApEn 0.20 u5min
		r = 0.20*sd, dos últimos 5 minutos
		ApEn 0.20 u10min
		r = 0.20*sd, dos últimos 10 minutos
		ApEn 0.20 u20min
		r = 0.20*sd, dos últimos 20 minutos
	Entropia Amostral	SampEn 0.10 50min
		r = 0.10*sd, dos últimos 50 minutos
		SampEn 0.15 50min
		r = 0.15*sd, dos últimos 50 minutos
		SampEn 0.20 50min
		r = 0.20*sd, dos últimos 50 minutos
		SampEn 0.10 p5min
		r = 0.10*sd, dos primeiros 5 minutos
		SampEn 0.10 p10min
		r = 0.10*sd, dos primeiros 10 minutos
		SampEn 0.10 p20min
		r = 0.10*sd, dos primeiros 20 minutos
		SampEn 0.15 p5min
		r = 0.15*sd, dos primeiros 5 minutos
		SampEn 0.15 p10min
		r = 0.15*sd, dos primeiros 10 minutos
		SampEn 0.15 p20min
		r = 0.15*sd, dos primeiros 20 minutos
		SampEn 0.20 p5min
		r = 0.20*sd, dos primeiros 5 minutos
		SampEn 0.20 p10min
		r = 0.20*sd, dos primeiros 10 minutos
		SampEn 0.20 p20min
		r = 0.20*sd, dos primeiros 20 minutos
		SampEn 0.10 u5min
		r = 0.10*sd, dos últimos 5 minutos
		SampEn 0.10 u10min
		r = 0.10*sd, dos últimos 10 minutos
		SampEn 0.10 u20min
		r = 0.10*sd, dos últimos 20 minutos
		SampEn 0.15 u5min
		r = 0.15*sd, dos últimos 5 minutos
		SampEn 0.15 u10min
		r = 0.15*sd, dos últimos 10 minutos
		SampEn 0.15 u20min
		r = 0.15*sd, dos últimos 20 minutos
		SampEn 0.20 u5min
		r = 0.20*sd, dos últimos 5 minutos
		SampEn 0.20 u10min
		r = 0.20*sd, dos últimos 10 minutos
		SampEn 0.20 u20min
		r = 0.20*sd, dos últimos 20 minutos
	Compressão	Brotli 50min
		Compressor Brotli, dos últimos 50 minutos
		Brotli p5min
		Compressor Brotli, dos primeiros 5 minutos
		Brotli p10min
		Compressor Brotli, dos primeiros 10 minutos

		Brotli p20min	Compressor Brotli, dos primeiros 20 minutos
		Brotli u5min	Compressor Brotli, dos últimos 5 minutos
		Brotli u10min	Compressor Brotli, dos últimos 10 minutos
		Brotli u20min	Compressor Brotli, dos últimos 20 minutos
		Brotli p5min	Compressor Brotli, dos primeiros 5 minutos
		Bzip 50min	Compressor Bzip, dos últimos 50 minutos
		Bzip p5min	Compressor Bzip, dos primeiros 5 minutos
		Bzip p10min	Compressor Bzip, dos primeiros 10 minutos
		Bzip p20min	Compressor Bzip, dos primeiros 20 minutos
		Bzip u5min	Compressor Bzip, dos últimos 5 minutos
		Bzip u10min	Compressor Bzip, dos últimos 10 minutos
		Bzip u20min	Compressor Bzip, dos últimos 20 minutos
		Gzip 50min	Compressor Gzip, dos últimos 50 minutos
		Gzip p5min	Compressor Gzip, dos primeiros 5 minutos
		Gzip p10min	Compressor Gzip, dos primeiros 10 minutos
		Gzip p20min	Compressor Gzip, dos primeiros 20 minutos
		Gzip u5min	Compressor Gzip, dos últimos 5 minutos
		Gzip u10min	Compressor Gzip, dos últimos 10 minutos
		Gzip u20min	Compressor Gzip, dos últimos 20 minutos
		Lzma	Compressor Lzma, dos últimos 50 minutos
		Lzma p5min	Compressor Lzma, dos primeiros 5 minutos
		Lzma p10min	Compressor Lzma, dos primeiros 10 minutos
		Lzma p20min	Compressor Lzma, dos primeiros 20 minutos
		Lzma u5min	Compressor Lzma, dos últimos 5 minutos
		Lzma u10min	Compressor Lzma, dos últimos 10 minutos
		Lzma u20min	Compressor Lzma, dos últimos 20 minutos
		Paq8l 50min	Compressor Paq8l, dos últimos 50 minutos
		Paq8l p5min	Compressor Paq8l, dos primeiros 5 minutos
		Paq8l p10min	Compressor Paq8l, dos primeiros 10 minutos
		Paq8l p20min	Compressor Paq8l, dos primeiros 20 minutos
		Paq8l u5min	Compressor Paq8l, dos últimos 5 minutos
		Paq8l u10min	Compressor Paq8l, dos últimos 10 minutos
		Paq8l u20min	Compressor Paq8l, dos últimos 20 minutos
		Ppmd 50min	Compressor Ppmd, dos últimos 50 minutos
		Ppmd p5min	Compressor Ppmd, dos primeiros 5 minutos
		Ppmd p10min	Compressor Ppmd, dos primeiros 10 minutos
		Ppmd p20min	Compressor Ppmd, dos primeiros 20 minutos
		Ppmd u5min	Compressor Ppmd, dos últimos 5 minutos
		Ppmd u10min	Compressor Ppmd, dos últimos 10 minutos
		Ppmd u20min	Compressor Ppmd, dos últimos 20 minutos
Índices lineares	Domínio do tempo	FHR_Base_Line	Linha de base da FCF
		FHR_Basal_Line	Linha basal da FCF
		Accelerations_Last_60_Min	Acelerações nos últimos 60 minutos
		Contractions_Last_60_Min	Contrações nos últimos 60 minutos
		Abnormal_STV_Last_60_Min	STV anormal nos últimos 60 minutos
		Average_STV_Last_60_Min	STV média nos últimos 60 minutos
		Abnormal_LTV_Last_60_Min	LTV anormal nos últimos 60 minutos

		Average_LTV_Last_60_Min	LTV média nos últimos 60 minutos
		Saltatory_Index_Last_60_Min	Índice Saltatório nos últimos 60 minutos
		Mild_Decelerations_Last_60_Min	Desacelerações moderadas nos últimos 60 minutos
	Domínio da frequência	LF	Frequência baixa
		VLF	Frequência muito baixa
		HF	Frequência alta
		LF/HF	Rácio: frequência baixa/frequência alta

2.3. Métodos estatísticos

2.3.1. Análise de dados

Os dados foram disponibilizados em Microsoft Excel® e foram analisados utilizando o software estatístico SPSS® versão 25.0. As variáveis categóricas e as ordinais são descritas pelas respectivas frequências absolutas e relativas, $n(\%)$. As variáveis contínuas são descritas pela média e respetivo desvio-padrão, $\bar{x} \pm sd$. A normalidade das variáveis contínuas é averiguada por visualização dos respetivos histogramas. Para comparar duas variáveis contínuas normalmente distribuídas utiliza-se o teste t de Student. No caso de variáveis contínuas não normais, é utilizado o teste não-paramétrico de Mann-Whitney. Para determinar a correlação entre duas variáveis contínuas normalmente distribuídas utiliza-se o coeficiente de correlação de Pearson, caso contrário aplica-se o coeficiente de correlação de Spearman.

Para a definição de acidemia, considera-se o pH da artéria umbilical (pHart): se $\text{pHart} \leq 7.15$, o caso pertence ao grupo dos acidémicos; caso contrário, pertence ao grupo dos normais (não acidémicos). A variável dicotómica resultante (acidémico *versus* não acidémico), constitui a variável dependente na análise discriminante. Como variáveis independentes são considerados os índices lineares do domínio da frequência, os índices lineares do domínio do tempo calculados pelo SisPorto® e os índices não lineares baseados em entropia e compressão. Valores de $p \leq 0.05$ são considerados significativos.

2.3.2. Pré-processamento

Cada um dos casos tem associado o registo dos batimentos cardíacos fetais que, geralmente, nos minutos finais que antecedem o parto, são fortemente contaminados por perdas de sinal e artefactos, devido aos movimentos fetais e maternos. Os dados utilizados

nesta dissertação foram previamente processados para outro estudo ainda não publicado. Foi implementado um algoritmo de pré-processamento introduzido em estudos realizados anteriormente [23]. Este algoritmo de pré-processamento do sinal deteta batimentos cardíacos fetais inferiores a 60 bpm, acima de 200 bpm e diferenças consecutivas superiores a 25 bpm. Esses batimentos são eliminados e substituídos por interpolação linear quando os períodos de perda de sinal não excedem os 2 segundos. Já no caso de períodos mais longos, um segmento anterior com igual comprimento e sem perda de sinal é replicado [22]. Este procedimento foi necessário para permitir o cálculo dos índices lineares do domínio da frequência e não lineares, de forma a reduzir contaminação por perdas de sinal e artefactos. Após este pré-processamento, todas as amostras foram arredondadas à unidade e foram descartados os últimos 5 minutos de cada traçado. Os últimos 50 minutos foram então seleccionados para realizar o estudo.

2.3. Classificação

De forma a avaliar e comparar a performance dos modelos encontrados pela AD, foram aplicadas duas abordagens de classificação. Da amostra disponível para este estudo, foram constituídos um conjunto de treino e um conjunto de teste. No conjunto de treino foi utilizada a técnica de validação cruzada *leave-one-out*, disponível no SPSS®, que classifica cada caso do conjunto de treino usando os restantes elementos do mesmo conjunto para a construção do modelo. Desta forma, é possível avaliar a classificação de um modelo em dados que não foram usados para o estimar. Adicionalmente, o modelo encontrado utilizando a totalidade dos elementos do conjunto de treino foi utilizado para classificar os elementos do conjunto de teste. Em ambos os casos, recorreu-se à percentagem de acerto total para a comparação entre os resultados.

3. Resultados

Do total de variáveis disponíveis para o estudo, os índices referentes aos movimentos dos fetos nos últimos 60 minutos, às desacelerações FIGO prolongadas e muito prolongadas dos últimos 60 minutos e às desacelerações repetidas nos últimos 20 minutos, foram retirados, uma vez que eram variáveis categóricas, resultando em 99 variáveis elegíveis para análise (Tabela 2). Em relação aos casos, foram selecionados aleatoriamente⁷ elementos para constituir dois conjuntos disjuntos de casos: um conjunto de treino e um conjunto de teste. De forma a cumprir o pressuposto de grupos equilibrados, e tendo em conta que a amostra possui apenas 20 casos de acidemia, foi constituído um conjunto de treino com 17 casos não acidémicos e 17 casos acidémicos e um conjunto de teste com 3 casos não acidémicos e 3 casos acidémicos. As características dos conjuntos podem ser consultadas na Tabela 3.

Tabela 3 - Características biométricas do conjunto de treino e do conjunto de teste. Fonte: elaboração própria.

Caraterísticas		Conjunto de treino	Conjunto de teste
Sexo, n(%)	Feminino	17 (50)	3 (50)
	Masculino	17 (50)	3 (50)
Tipo de parto, n(%)	Normal	16 (47.1)	1 (16.7)
	Cesariana	11 (32.4)	3 (50)
	Instrumental	7 (20.6)	2 (33.3)
pH arteria umbilical, n(%)	≤ 7.15	17 (50)	3 (50)
	> 7.15	17 (50)	3 (50)
Idade gestacional, n(%)	36	-	1 (16.7)
	37	3 (8.8)	-
	38	6 (17.6)	1 (16.7)
	39	13 (38.2)	2 (33.3)
	40	6 (17.6)	2 (33.3)
	41	6 (17.6)	-
Índice de Apgar, $\bar{x} \pm sd$	1º min	7.97 ± 1.314	7 ± 1.673
	5º min	9.41 ± 0.783	8.97 ± 1.211

⁷ Através da função *aleatório* do Microsoft Excel®.

	10º min	9.78 ± 0.518	9.50 ± 0.837
Peso (gramas), $\bar{x} \pm sd$		3166.68 ± 441.972	3096 ± 359.155
Duração dos traçados (min), $\bar{x} \pm sd$	Mínimo: 60 Máximo: 780	210.82 ± 138.274	235 ± 179.206

Seguidamente foi aplicado o teste não paramétrico de Mann-Whitney (uma vez que os histogramas das variáveis não corresponderam a distribuições normais) para duas amostras independentes ao conjunto de treino, com o objetivo de testar a hipótese nula de que as distribuições das variáveis na população são as mesmas entre os grupos dos académicos e dos não académicos. Com base nos resultados deste teste foram, posteriormente, selecionadas as 32 variáveis que se mostraram estatisticamente significativas (Tabela 4).

Tabela 4 - Resultados do teste para igualdade das distribuições no conjunto de treino - Teste Mann-Whitney. Apenas as variáveis significativas são apresentadas. Fonte: elaboração própria.

Variáveis	U e Mann-Whitney	Wilcoxon W	Z	p-value
ApEn 0.10 50min	53.000	206.000	-3.152	0.002
Low Frequency	58.000	211.000	-2.979	0.003
SampEn 0.10 50min	60.000	213.000	-2.910	0.004
Very Low Frequency	61.000	214.000	-2.876	0.004
ApEn 0.20 50min	67.000	220.000	-2.669	0.008
Ppm p5min	67.000	220.000	-2.669	0.008
ApEn 0.10 u20min	68.000	221.000	-2.635	0.008
ApEn 0.15 50min	69.000	222.000	-2.600	0.009
Bzip p5min	71.000	224.000	-2.532	0.011
SampEn 0.15 50min	72.000	225.000	-2.497	0.013
ApEn 0.15 u20min	73.000	226.000	-2.463	0.014
SampEn 0.15 u20min	73.000	226.000	-2.463	0.014
SampEn 0.10 u20min	74.000	227.000	-2.428	0.015
ApEn 0.10 u10min	75.000	228.000	-2.394	0.017
Gzip p5min	75.000	228.000	-2.394	0.017
SampEn 0.20 50min	79.000	232.000	-2.256	0.024
ApEn 0.10 p20min	80.000	233.000	-2.222	0.026
SampEn 0.20 u20min	81.000	234.000	-2.187	0.029
ApEn 0.10 p10min	82.000	235.000	-2.153	0.031
ApEn 0.10 u5min	82.000	235.000	-2.153	0.031
Lzma p5min	83.000	236.000	-2.118	0.034
Paq8l p5min	84.000	237.000	-2.084	0.037
ApEn 0.15 p20min	84.000	237.000	-2.084	0.037
SampEn 0.10 p20min	85.000	238.000	-2.049	0.040
Gzip p10min	85.000	238.000	-2.049	0.040
ApEn 0.15 p5min	86.000	239.000	-2.015	0.044
ApEn 0.15 u10min	86.000	239.000	-2.015	0.044
Bzip p10min	86.500	239.500	-1.998	0.046

High frequency	87.000	240.000	-1.981	0.048
Raio LF /HF	87.000	240.000	-1.981	0.048
ApEn 0.20 u20min	87.000	240.000	-1.981	0.048
Ppm p10min	87.000	240.000	-1.981	0.048

3.1. Análise discriminante aplicada ao conjunto de treino

Foram selecionadas todas as variáveis significativas na discriminação entre os dois grupos no conjunto de treino apresentadas na Tabela 4 e, calculadas as correlações de Spearman entre todos os pares de variáveis (Tabela 26 em Anexos). Tendo em conta os pressupostos de pelo menos 5k casos e uma vez que existem 34 elementos no conjunto de treino, foram selecionadas k=5 variáveis. De forma a que as correlações entre si fossem as mais baixas e selecionando índices lineares do domínio do tempo, do domínio da frequência e índices não lineares, foram escolhidos os seguintes conjuntos de variáveis independentes a incorporar no modelo de análise discriminante:

Primeiro conjunto de variáveis: Gzip p5min, SampEn 0.10 p20min, ApEn 0.10 50min, LF e o Rácio LF/HF.

Segundo conjunto de variáveis: SampEn 0.10 50min, LF, Pmpd p5min, Rácio LF/HF e ApEn 0.10 p20min.

3.1.1. Método das *variáveis independentes juntas* aplicado ao primeiro conjunto de variáveis

Para o primeiro modelo foi realizada a AD recorrendo ao *método das variáveis independentes juntas*, considerando os 34 casos do conjunto de treino na análise e o primeiro conjunto de variáveis. A Tabela 5 indica a média e o desvio-padrão para cada variável em cada grupo (académicos *versus* não académicos) e no total.

Tabela 5 - Média e desvio-padrão das variáveis do primeiro modelo de AD no grupo dos não académicos, no grupo dos académicos e nos dois em conjunto. Fonte: elaboração própria.

Variável		$\bar{x} \pm sd$
Não académicos	Gzip p5min	12.34±2.39
	SampEn 0.10 p20min	0.64±0.27
	ApEn 0.10 50min	0.74±0.28

	Low Frequency	12.32±6.60
	Rácio LF /HF	5.01±1.61
Académicos	Gzip p5min	14.36±2.35
	SampEn 0.10 p20min	0.45±0.21
	ApEn 0.10 50min	0.45±0.17
	Low Frequency	6.02±3.28
	Rácio LF /HF	4.15±1.93
Total	Gzip p5min	13.35±2.54
	SampEn 0.10 p20min	0.54±0.25
	ApEn 0.10 50min	0.60±0.27
	Low Frequency	9.17±6.05
	Rácio LF /HF	4.58±1.80

Para o teste da igualdade das médias de cada variável entre os grupos académicos e não académicos, foi utilizado o Lambda de Wilks, com resultados apresentados na Tabela 6. A estatística Lambda de Wilks testa se as médias são significativamente diferentes de forma a diferenciar os grupos. Podemos verificar que tal não se verifica apenas na variável “Rácio LF/HF” ($p\text{-value} = 0.170$), o que significa que a variável não distingue significativamente o grupo de académicos do grupo de não académicos. De qualquer forma, nenhum dos Lambdas de Wilks calculados estão próximos de zero, sendo o índice ApEn 0.10 50min o que melhor discrimina, isoladamente, os dois grupos definidos pelo pH.

Tabela 6 - Lambda de Wilks das variáveis no modelo de AD e teste de igualdade das médias. Fonte: elaboração própria.

Variáveis	Lambda de Wilks	Z	p-value
Gzip p5min	0.84	6.17	0.020
SampEn 0.10 p20min	0.86	5.40	0.030
ApEn 0.10 50min	0.69	14.11	< 0.001
Low Frequency	0.72	12.44	< 0.001
Rácio LF/HF	0.94	1.97	0.170

A Tabela 7 mostra a matriz de covariância agrupada e a matriz de correlação agrupada. Esta tabela contribui para a avaliação da relação entre as variáveis. É possível verificar que a maior correlação acontece entre a Low Frequency e a ApEn 0.10 50min.

Tabela 7 - Matriz de covariância e de correlação. Fonte: elaboração própria.

Variáveis		Gzip p5min	SampEn 0.10 p20min	ApEn 0.10 50min	Low Frequency	Rácio LF /HF
Covariância	Gzip p5min	5.59	-0.08	0.13	1.87	-0.40
	SampEn 0.10 p20min	-0.08	0.06	0.02	0.63	-0.01
	ApEn 0.10 50min	0.13	0.02	0.05	0.66	-0.03
	Low Frequency	1.87	0.63	0.66	27.16	0.55
	Rácio LF /HF	-0.40	-0.01	-0.03	0.55	3.14
Correlação	Gzip p5min	1.00	-0.15	0.25	0.15	-0.10
	SampEn 0.10 p20min	-0.15	1.00	0.45	0.50	-0.02
	ApEn 0.10 50min	0.25	0.45	1.00	0.56	-0.09
	Low Frequency	0.15	0.50	0.56	1.00	0.06
	Rácio LF /HF	-0.10	-0.02	-0.09	0.06	1.00

No resultado seguinte (Tabela 8) é possível ver a matriz de covariância no grupo dos académicos e dos não académicos, de forma a verificar a homogeneidade de covariância entre os grupos. Aplicando o teste M'Box, que testa se as matrizes de covariância entre os dois grupos são iguais, rejeita-se a hipótese nula ($p\text{-value} = 0.020$). O facto de não ser possível assumir homogeneidade das matrizes de covariância não impede que se prossiga com a AD neste grupo de variáveis, mas pode representar uma limitação do modelo.

Tabela 8 - Matriz de covariâncias em cada grupo do conjunto de treino. Fonte: elaboração própria.

Variáveis		Gzip p5min	SampEn 0.10 p20min	ApEn 0.10 50min	Low Frequency	Rácio LF /HF
Não académicos	Gzip p5min	5.69	-0.02	0.08	1.50	-0.17
	SampEn 0.10 p20min	-0.02	0.07	0.04	1.21	0.08
	ApEn 0.10 50min	0.08	0.04	0.08	0.90	0.05
	Low Frequency	1.50	1.21	0.90	43.58	1.70
	Rácio LF /HF	-0.17	0.08	0.05	1.70	2.58
Académicos	Gzip p5min	5.50	-0.15	0.18	2.25	-0.64
	SampEn 0.10 p20min	-0.15	0.04	0.01	0.05	-0.10
	ApEn 0.10 50min	0.18	0.01	0.03	0.43	-0.12
	Low Frequency	2.25	0.05	0.43	10.75	-0.60
	Rácio LF /HF	-0.64	-0.10	-0.12	-0.60	3.71
Total	Gzip p5min	6.47	-0.18	-0.02	-1.46	-0.83
	SampEn 0.10 p20min	-0.18	0.06	0.04	0.92	0.03
	ApEn 0.10 50min	-0.02	0.04	0.07	1.12	0.03
	Low Frequency	-1.46	0.92	1.12	36.58	1.92

	Rácio LF /HF	-0.83	0.03	0.03	1.92	3.24
--	--------------	-------	------	------	------	------

O número de valores próprios (Tabela 9) será igual ao número de funções que a análise irá produzir. Contudo, este número depende da quantidade de grupos em estudo, uma vez que é definido como sendo o nº de grupos menos 1 unidade. Assim, neste caso, como a variável dependente tem apenas duas categorias, a análise irá produzir apenas uma função. O valor próprio e a correlação canónica associados a esta função podem ser consultados na tabela seguinte.

Tabela 9 - Valor próprio. Fonte: elaboração própria.

Função	Valor próprio	% de variância	% cumulativa	Correlação canónica
1	1.006	100.0	100.0	0.71

O valor próprio da função discriminante é pouco superior a 1 (1.006) e corresponde a 100% da variância explicada no que diz respeito às diferenças entre os grupos. A correlação canónica entre a função discriminante e os grupos é de 0.71, o que indica que o *effect size* (efeito das variáveis independentes na variável a prever) do modelo encontrado é 0.50 (0.71^2).

No entanto, pelo teste Lambda de Wilks (Tabela 10) aplicado à função discriminante obtida, verifica-se que a função consegue diferenciar os grupos de forma significativa (*p-value* = 0.001), o que indica que o modelo encontrado é forte.

Tabela 10 - Lambda de Wilks da função discriminante do modelo de AD e teste de igualdade das médias. Fonte: elaboração própria.

Teste de funções	Lambda de Wilks	Qui-quadrado	df	p-value
1	0.50	20.53	5	0.001

Os valores dos coeficientes padronizados da função de discriminante (Tabela 11) podem ser utilizados para avaliar a importância relativa de cada variável explicativa para a função discriminante. Verifica-se que as variáveis ApEn 0.10 50min e Gzip p5min são as variáveis com maior peso no modelo encontrado (0.71 e -0.69, respetivamente). A matriz de estrutura (Tabela 11) apresenta as correlações entre as variáveis na função discriminante

e a função discriminante canónica e por isso auxilia também na interpretação da contribuição de cada variável na função discriminante. Neste caso, as correlações mais elevadas verificam-se para as os índices ApEn 0.10 50min (0.66) e Low Frequency (0.62) presentes no modelo.

Tabela 11 - Coeficientes padronizados da função discriminantes e matriz de estrutura. Fonte: elaboração própria.

Variáveis	Coeficientes	Matriz de estrutura
Gzip p5min	-0.69	-0.44
SampEn 0.10 p20min	-0.21	0.41
ApEn 0.10 50min	0.71	0.66
Low Frequency	0.42	0.62
Rácio LF /HF	0.21	0.25

A Tabela 12 mostra os valores dos coeficientes não padronizados da função discriminante, que dá a direção encontrada pela técnica de análise discriminante para projetar os dados de forma a melhor separar os dois grupos.

Tabela 12 - Coeficientes da função discriminante de Fisher. Fonte: elaboração própria.

Variáveis	Função 1
Gzip p5min	-0.292
SampEn 0.10 p20min	-0.900
ApEn 0.10 50min	3.123
Low Frequency	0.081
Rácio LF /HF	0.120
(Constante)	1.223
Coeficientes não padronizados	

Com base na Tabela 12, é possível escrever a função discriminante de Fisher para este primeiro modelo:

$$Z = 1.223 - 0.292*(Gzip p5min) - 0.900*(SampEn 0.10 p20min) + \\ +3.123*(ApEn 0.10 50min) + 0.081*(Low Frequency) + 0.120*(Rácio LF/HF)$$

Na Tabela 13 estão representados os coeficientes das funções score discriminante para cada grupo, que não têm qualquer interpretação discriminativa. É através das duas funções obtidas com estes coeficientes, que se classificam os novos elementos comparando o score obtido em cada uma delas. Representa-se o score discriminante na função dos não académicos e na função dos académicos por $Z_{\text{não académicos}}$ e $Z_{\text{académicos}}$, respetivamente. O grupo escolhido para classificar um caso novo será o grupo que maximiza a função score discriminante.

Tabela 13 - Coeficientes das funções score de classificação. Fonte: elaboração própria.

	Não académicos	Académicos
Gzip p5min	2.56	3.12
SampEn 0.10 p20min	15.92	17.67
ApEn 0.10 50min	4.99	-1.08
Low Frequency	-0.25	-0.41
Rácio LF /HF	2.07	1.84
(Constante)	-27.04	-29.42
Funções discriminantes lineares de Fisher		

$$Z_{\text{não académicos}} = -27.04 + 2.56*(\text{Gzip p5min}) + 15.92*(\text{SampEn 0.10 p20min}) + 4.99*$$

$$*(\text{ApEn 0.10 50min}) - 0.25*(\text{Low Frequency}) + 2.07*(\text{Rácio LF/HF})$$

$$Z_{\text{académicos}} = -29.42 + 3.12*(\text{Gzip p5min}) + 17.67*(\text{SampEn 0.10 p20min}) - 1.08*$$

$$*(\text{ApEn 0.10 50min}) - 0.41*(\text{Low Frequency}) + 1.84*(\text{Rácio LF/HF})$$

Por último, obtém-se a Tabela 14, que mostra a efetividade do modelo, isto é, qual a percentagem de acerto que o modelo obteve nas classificações dos casos usados para o treino. Neste caso, o modelo conseguiu acertar a classificação em 85.3% dos casos. Para que seja possível considerar que o modelo tem uma classificação razoável, no mínimo, é necessário que esta percentagem de acerto seja maior que a probabilidade de o caso ser classificado em cada categoria mais 25%. Neste caso, a probabilidade de ser classificado num dos dois grupos é 50%, logo $50\% + 25\% = 75\% < 85.3\%$ ou seja, trata-se de uma classificação razoável. Devido ao reduzido tamanho de casos na amostra disponível é usada a técnica de validação cruzada *leave-one-out* para avaliação da performance do modelo. Na Tabela 14 é possível verificar a performance do modelo nesta classificação que resultou em 27 (79.4%) casos corretamente classificados.

Tabela 14 - Resultados de classificação original e com validação cruzada. Fonte: elaboração própria.

			Associação ao grupo previsto		Total
			Não académicos	Académicos	
Original ^a	Contagem	Não académicos	14	3	17
		Académicos	2	15	17
	%	Não académicos	82.4	17.6	100
		Académicos	11.8	88.2	100
Com validação cruzada ^b	Contagem	Não académicos	14	3	17
		Académicos	4	13	17
	%	Não académicos	82.4	17.6	100
		Académicos	23.5	76.5	100

a. 85.3% de casos agrupados originais foram bem classificados.

b. 79.4% dos casos com validação cruzada foram bem classificados.

3.1.2. Método *Stepwise* aplicado ao primeiro conjunto de variáveis

Para o segundo modelo foi realizada a AD, aplicada novamente ao primeiro conjunto de variáveis, mas recorrendo ao método *Stepwise*, considerando os 34 casos do conjunto de treino na análise. A Tabela 15 indica as variáveis incluídas no modelo recorrendo ao Lambda de Wilks: em cada passo, a variável que minimiza o Lambda de Wilks total é inserida, até que não ocorram variações significativas do Lambda de Wilks. A variável ApEn 0.10 50min foi incluída no passo 1, já no passo 2 foi incluída a Gzip p5min.

Tabela 15 - Variáveis incluídas/excluídas^{a,b,c,d} no modelo recorrendo ao Lambda de Wilks. Fonte: elaboração própria.

Passo	Inseridas	Lambda de Wilks							
		Estatística	df1	df2	df3	F exato			
						Estatística	df1	df2	p-value
1	ApEn								
	0.10	0.694	1	1	32.000	14.107	1	32.000	0.001
	50min								
2	Gzip p5min	0.547	2	1	32.000	12.857	2	31.000	< 0.001

Em cada passo, a variável que minimiza o Lambda de Wilks geral é inserida.

a. O número máximo de etapas é 10.

b. O F parcial mínimo a ser inserido é 3.84.

c. O F parcial máximo a ser removido é 2.71.

d. Nível f, tolerância ou VIN insuficiente para cálculos adicionais.

A Tabela 16 apresenta as variáveis consideradas discriminativas em cada passo juntamente com os valores de tolerância (T) associados. Os valores de tolerância são todos superiores a 0.9 o que pode indicar a existência de colinearidade entre as variáveis selecionadas [18].

Tabela 16 - Variáveis incluídas na análise pelo método *Stepwise*. Fonte: elaboração própria.

Passo		Tolerância (T)	F a ser removido	Lambda de Wilks
1	ApEn 0.10 50min	1.000	14.107	
2	ApEn 0.10 50min	0.938	16.543	0.838
	Gzip p5min	0.938	8.362	0.694

Por outro lado as variáveis excluídas da análise foram a SampEn 0.10 p20min, Low Frequency e o rácio LF/HF (Tabela 17).

Tabela 17 - Variáveis excluídas da análise pelo método *Stepwise*. Fonte: elaboração própria.

Passo		Tolerância (T)	Mín. Tolerância	F a ser inserido	Lambda de Wilks
0	Gzip p5min	1.000	1.000	6.173	0.838
	SampEn 0.10 p20min	1.000	1.000	5.402	0.856
	ApEn 0.10 50min	1.000	1.000	14.107	0.694
	Low Frequency	1.000	1.000	12.438	0.720
	Racio LF /HF	1.000	1.000	1.974	0.942
1	Gzip p5min	0.938	0.938	8.362	0.547
	SampEn 0.10 p20min	0.802	0.802	0.357	0.686
	Low Frequency	0.686	0.686	1.979	0.652
	Racio LF /HF	0.992	0.992	2.029	0.651
	SampEn 0.10 p20min	0.732	0.702	0.056	0.546
2	Low Frequency	0.685	0.658	1.610	0.519
	Racio LF /HF	0.987	0.933	1.106	0.527

O valor próprio da função discriminante é inferior a 1 (0.830) e corresponde, tal como no modelo anterior, a 100% da variância explicada no que diz respeito às diferenças entre os grupos (Tabela 18). A correlação canónica entre a função discriminante e os grupos é de 0.673, o que indica que o *effect size* (efeito das variáveis independentes na variável a prever) do modelo encontrado é 0.45 (0.673²).

Tabela 18 - Valor próprio. Fonte: elaboração própria.

Função	Valor próprio	% de variância	% cumulativa	Correlação canónica
--------	---------------	----------------	--------------	---------------------

1	.830	100	100.0	0.673
---	------	-----	-------	-------

Por outro lado, o resultado do teste Lambda de Wilks (Tabela 19) indica que a função discriminante obtida consegue diferenciar os grupos de forma estatisticamente significativa ($p\text{-value} < 0.001$) e que, portanto, o modelo é forte.

Tabela 19 - Lambda Wilks. Fonte: elaboração própria.

Teste de funções	Lambda de Wilks	Qui-quadrado	df	p-value
1	0.547	18.725	2	< 0.001

Na Tabela 20 verifica-se que a variável com maior peso na função discriminante é a ApEn 0.10 50min (0.904). Neste caso, verifica-se que a correlação mais elevada com a função discriminante canónica ocorre também para a mesma variável (0.729).

Tabela 20 - Coeficientes padronizados da função discriminantes e matriz de estrutura. Fonte: elaboração própria.

Variáveis	Coeficientes	Matriz de estrutura
Gzip p5min	-0.707	-0.482
SampEn 0.10 p20min ^a	-	0.506
ApEn 0.10 50min	0.904	0.729
Low Frequency ^a	-	0.400
Rácio LF /HF ^a	-	-0.011
a. Variável não utilizada na análise.		

A Tabela 21 mostra os valores das funções de Fisher não padronizadas e é através dos valores obtidos nesta tabela que se escreve a função discriminante.

Tabela 21 - Coeficientes da função de Fisher. Fonte: elaboração própria.

Variáveis	Função 1
Gzip p5min	-0.299
ApEn 0.10 50min	3.981
(Constante)	1.614
Coeficientes não padronizados	

Assim, a função discriminante de Fisher para o segundo modelo é dada por:

$$Z = 1.614 - 0.299*(Gzip\ p5min) + 3.981*(ApEn\ 0.10\ 50min)$$

Na Tabela 22 estão representados os coeficientes das funções score discriminante para cada grupo. É através destes coeficientes que se classificam os novos elementos.

Tabela 22 - Coeficientes das funções de classificação. Fonte: elaboração própria.

	Não académicos	Académicos
Gzip p5min	1.986	2.513
ApEn 0.10 50min	9.262	2.227
(Constante)	-16.385	-19.238
Funções discriminantes lineares de Fisher		

Funções score discriminante de Fisher:

$$Z_{\text{não académicos}} = -16.385 + 1.986*(Gzip\ p5min) + 9.262*(ApEn\ 0.10\ 50min)$$

$$Z_{\text{académicos}} = -19.238 + 2.513*(Gzip\ p5min) + 2.227*(ApEn\ 0.10\ 50min)$$

Por último, obtém-se a Tabela 23, que mostra as classificações com o método *Stepwise*. Neste caso, o modelo conseguiu acertar a classificação em 85.3% dos casos, o que representa uma classificação razoável. Na Tabela 23 é ainda possível verificar a performance do modelo pela técnica de validação cruzada *leave-one-out*, que resultou em 27 (85.3%) casos corretamente classificados.

Tabela 23 - Resultados de classificação original e com validação cruzada. Fonte: elaboração própria.

			Associação ao grupo previsto		Total
			Não académicos	Académicos	
Original ^a	Contagem	Não académicos	14	3	17
		Académicos	2	15	17
	%	Não académicos	82.4	17.6	100
		Académicos	11.8	88.2	100
Com validação cruzada ^b	Contagem	Não académicos	14	3	17
		Académicos	2	15	17
	%	Não académicos	82.4	17.6	100
		Académicos	11.8	88.2	100

- a. 85.3% de casos agrupados originais foram bem classificados.
- b. 85.3% dos casos com validação cruzada foram bem classificados.

3.1.3. Método das *variáveis independentes juntas* aplicado ao segundo conjunto de variáveis

Neste segundo modelo foi realizada a AD recorrendo ao *método das variáveis independentes juntas*, considerando os 34 casos do conjunto de treino na análise e o segundo conjunto de variáveis. A Tabela 24 indica a média e o desvio-padrão para cada variável em cada grupo (académicos *versus* não académicos) e no total.

Tabela 24 - Média e desvio-padrão das variáveis do segundo modelo de AD no grupo dos não académicos, no grupo dos académicos e nos dois em conjunto. Fonte: elaboração própria.

	Variável	$\bar{x} \pm sd$
Não académicos	SampEn 0.10 50min	0.52±0.23
	Low Frequency	12.32±6.60
	Ppmd p5min	10.10±2.13
	Rácio LF /HF	5.01±1.61
	ApEn 0.10 p20min	0.83±0.24
Académicos	SampEn 0.10 50min	0.31±0.15
	Low Frequency	6.02±3.28
	Ppmd p5min	12.12±2.25
	Rácio LF /HF	4.15±1.93
	ApEn 0.10 p20min	0.62±0.25
Total	SampEn 0.10 50min	0.41±0.22
	Low Frequency	9.17±6.05
	Ppmd p5min	11.11±2.39
	Rácio LF /HF	4.58±1.80
	ApEn 0.10 p20min	0.72±0.26

Na Tabela 25 estão representados os valores para o Lambda de Wilks. Tal como aconteceu no primeiro conjunto de variáveis o “Rácio LF/HF” ($p\text{-value} = 0.170$), é a única variável neste conjunto que não distingue significativamente o grupo de académicos do grupo de não académicos. De qualquer forma, nenhum dos Lambdas de Wilks calculados

estão próximos de zero, nesta situação é possível afirmar que a variável que melhor discrimina os dois grupos (acadêmicos e não acadêmicos) é a Low Frequency.

Tabela 25 - Lambda de Wilks das variáveis no modelo de AD e teste de igualdade das médias. Fonte: elaboração própria.

Variáveis	Lambda de Wilks	Z	p-value
SampEn 0.10 50min	0.76	10.12	0.003
Low Frequency	0.72	12.44	0.001
Ppmd p5min	0.82	7.23	0.011
Rácio LF /HF	0.94	1.97	0.170
ApEn 0.10 p20min	0.83	6.54	0.015

A Tabela 26 mostra a matriz de covariância agrupada e a matriz de correlação agrupada. Esta tabela contribui para a avaliação da relação entre as variáveis. É possível verificar que a maior correlação acontece entre a Low Frequency e a SampEn 0.10 50 min.

Tabela 26 - Matriz de covariância e de correlação. Fonte: elaboração própria.

		SampEn 0.10 50min	Low Frequency	Ppmd p5min	Rácio LF /HF	ApEn 0.10 p20min
Covariância	SampEn 0.10 50min	0.038	0.569	0.059	-0.013	0.020
	Low Frequency	0.569	27.163	1.552	0.546	0.502
	Ppmd p5min	0.059	1.552	4.799	-0.268	-0.188
	Rácio LF /HF	-0.013	0.546	-0.268	3.145	0.002
	ApEn 0.10 p20min	0.020	0.502	-0.188	0.002	0.059
Correlação	SampEn 0.10 50min	1.000	0.563	0.140	-0.039	0.432
	Low Frequency	0.563	1.000	0.136	0.059	0.396
	Ppmd p5min	0.140	0.136	1.000	-0.069	-0.352
	Rácio LF /HF	-0.039	0.059	-0.069	1.000	0.004
	ApEn 0.10 p20min	0.432	0.396	-0.352	0.004	1.000

Atráves da Tabela 27 é possível verificar a homogeneidade da matriz de covariância entre os grupos. Com o teste M'Box, que testa se as matrizes de covariância entre os dois grupos são iguais, não se rejeita a hipótese nula ($p\text{-value} = 0.065$). O que indica que se pode assumir homogeneidade das matrizes de covariância.

Tabela 27 - Matriz de covariâncias em cada grupo do conjunto de treino. Fonte: elaboração própria.

	Variáveis	SampEn 0.10 50min	Low Frequency	Ppmd p5min	Rácio LF /HF	ApEn 0.10 p20min
Não acadêmicos	SampEn 0.10 50min	0.052	0.765	-0.011	0.064	0.036
	Low Frequency	0.765	43.576	0.771	1.696	1.016
	Ppmd p5min	-0.011	0.771	4.517	-0.212	-0.057
	Rácio LF /HF	0.064	1.696	-0.212	2.579	0.108
	ApEn 0.10 p20min	0.036	1.016	-0.057	0.108	0.058
Acadêmicos	SampEn 0.10 50min	0.023	0.373	0.130	-0.090	0.005
	Low Frequency	0.373	10.749	2.332	-0.603	-0.011
	Ppmd p5min	0.130	2.332	5.082	-0.323	-0.319
	Rácio LF /HF	-0.090	-0.603	-0.323	3.710	-0.105
	ApEn 0.10 p20min	0.005	-0.011	-0.319	-0.105	0.061
Total	SampEn 0.10 50min	0.048	0.895	-0.053	0.034	0.031
	Low Frequency	0.895	36.578	-1.776	1.918	0.834
	Ppmd p5min	-0.053	-1.776	5.705	-0.704	-0.293
	Rácio LF /HF	0.034	1.918	-0.704	3.237	0.049
	ApEn 0.10 p20min	0.031	0.834	-0.293	0.049	0.069

A Tabela 28 apresenta o valor próprio da função discriminante que é 0.855. Já a correlação canônica entre a função discriminante e os grupos é de 0.679, o que indica que o *effect size* do modelo é 0.46 (0.679^2).

Tabela 28 - Valor próprio. Fonte: elaboração própria.

Função	Valor próprio	% de variância	% cumulativa	Correlação canônica
1	0.855	100.0	100.0	0.679

O teste Lambda de Wilks (Tabela 29) apresenta um *p-value* = 0.003, o que indica que esta função diferencia de um modo significativo os grupos em estudo.

Tabela 29 - Lambda de Wilks da função discriminante do modelo de AD e teste de igualdade das médias. Fonte: elaboração própria.

Teste de funções	Lambda de Wilks	Qui-quadrado	df	p-value
1	0.539	18.23	5	0.003

Verifica-se que as variáveis Low Frequency (0.57) e Ppmd p5min (-0.71) são as variáveis com maior peso no modelo em causa. A Tabela 31 apresenta a matriz de estrutura onde se verifica que as maiores correlações com a função discriminante ocorrem com as variáveis Low Frequency (0.67) e a SampEn 0.10 50min (0.61).

Tabela 30 - Coeficientes padronizados da função discriminantes e matriz de estrutura. Fonte: elaboração própria.

Variáveis	Coeficientes	Matriz de estrutura
SampEn 0.10 50min	0.48	0.61
Low Frequency	0.57	0.67
Ppmd p5min	-0.71	-0.51
Rácio LF /HF	0.21	0.27
ApEn 0.10 p20min	-0.19	0.49

Os valores dos coeficientes não padronizados da função discriminante apresentados na Tabela 31 indicam a direção encontrada pela técnica deAD que melhor separa os dois grupos definidos pela variável dependente.

Tabela 31 - Coeficientes da função discriminante de Fisher. Fonte: elaboração própria.

Variáveis	Função 1
SampEn 0.10 50min	2.483
Low Frequency	0.108
Ppmd p5min	-0.325
Rácio LF /HF	0.116
ApEn 0.10 p20min	-0.796
(Constante)	1.641
Coeficientes não padronizados	

Apartir da Tabela 31, é possível escrever a função discriminante de Fisher para este terceiro modelo:

$$Z = 1.641 + 2.483*(\text{SampEn 0.10 50min}) + 0.108*(\text{Low Frequency}) - 0.325*(\text{Ppmd p5min}) + 0.116*(\text{Rácio LF/HF}) - 0.796*(\text{ApEn 0.10 p20min})$$

Na Tabela 32 estão representados os coeficientes das funções score discriminante para cada grupo.

Tabela 32 - Coeficientes das funções score de classificação. Fonte: elaboração própria.

	Não académicos	Académicos
SampEn 0.10 50min	-2.17	-6.62
Low Frequency	-0.24	-0.44
Ppmd p5min	3.39	3.97
Rácio LF /HF	1.90	1.69
ApEn 0.10 p20min	27.46	28.89
(Constante)	-31.93	-34.88
Funções discriminantes lineares de Fisher		

Com as duas funções score discriminante obtidas com estes coeficientes é possível classificar os novos elementos:

$$Z_{\text{não académicos}} = -31.93 - 2.17 * (\text{SampEn 0.10 50min}) - 0.24 * (\text{Low Frequency}) + 3.39 * (\text{Ppmd p5min}) + 1.90 * (\text{Rácio LF/HF}) + 27.46 * (\text{ApEn 0.10 p20min})$$

$$Z_{\text{académicos}} = -34.88 - 6.62 * (\text{SampEn 0.10 50min}) - 0.44 * (\text{Low Frequency}) + 3.97 * (\text{Ppmd p5min}) + 1.69 * (\text{Rácio LF/HF}) + 28.89 * (\text{ApEn 0.10 p20min})$$

Por fim, obtém-se a Tabela 33, que apresenta a efetividade do modelo. Este segundo conjunto teve uma percentagem de acerto igual ao primeiro conjunto. Isto é, classificou corretamente 85.3% dos casos e na validação cruzada *leave-one-out*, 79.4%.

Tabela 33 - Resultados de classificação original e com validação cruzada. Fonte: elaboração própria.

			Associação ao grupo previsto		Total
			Não académicos	Académicos	
Original ^a	Contagem	Não académicos	14	3	17
		Académicos	2	15	17
	%	Não académicos	82.4	17.6	100
		Académicos	11.8	88.2	100
Com validação cruzada ^b	Contagem	Não académicos	13	4	17
		Académicos	3	14	17
	%	Não académicos	76.5	23.5	100
		Académicos	17.6	82.4	100

a. 85.3% de casos agrupados originais foram bem classificados.

b. 79.4% dos casos com validação cruzada foram bem classificados.

3.1.4. Método *Stepwise* aplicado ao segundo conjunto de variáveis

Aplicou-se novamente a AD ao segundo conjunto, mas desta vez o método utilizado foi o método *Stepwise*. A Tabela 34 apresenta as variáveis incluídas no modelo recorrendo ao Lambda de Wilks: a variável Low Frequency foi incluída no passo 1, já no passo 2 a variável incluída foi a Ppmd p5min.

Tabela 34 - Variáveis incluídas/excluídas^{a,b,c,d} no modelo recorrendo ao Lambda de Wilks. Fonte: elaboração própria.

Passo	Inseridas	Lambda de Wilks							
		Estatística	df1	df2	df3	F exato			
						Estatística	df1	df2	p-value
1	Low Frequency	0.720	1	1	32.000	12.438	1	32.000	0.001
2	Ppmd p5min	0.585	2	1	32.000	10.977	2	31.000	<0.001

Em cada passo, a variável que minimiza o Lambda de Wilks geral é inserida.

a. O número máximo de etapas é 10.

b. O F parcial mínimo a ser inserido é 3.84.

c. O F parcial máximo a ser removido é 2.71.

d. Nível f, tolerância ou VIN insuficiente para cálculos adicionais.

Na Tabela 35 estão representadas as variáveis consideradas discriminativas em cada passo e os respetivos valores de tolerância (T). Todos os valores de tolerância são superiores a 0.9, o que pode indicar a existência de colinearidade nas variáveis selecionadas [18].

Tabela 35 - Variáveis incluídas pelo método *Stepwise*. Fonte: elaboração própria.

Passo		Tolerância (T)	F a ser removido	Lambda de Wilks
1	Low Frequency	1.000	12.438	
2	Low Frequency	0.932	12.197	0.816
	Ppmd p5min	0.932	7.132	0.720

As variáveis excluídas deste modelo, como se pode ver na Tabela 36, foram a SampEn 0.10 p20min, o rácio LF/HF e a ApEn 0.10 p20min..

Tabela 36 - Variáveis excluídas pelo método *Stepwise*. Fonte: elaboração própria.

Passo		Tolerância (T)	Mín. Tolerância	F a ser inserido	Lambda de Wilks
0	SampEn 0.10 50min	1.000	1.000	10.120	0.760
	Low Frequency	1.000	1.000	12.438	0.720
	Ppmd p5min	1.000	1.000	7.229	0.816
	Rácio LF /HF	1.000	1.000	1.974	0.942
	ApEn 0.10 p20min	1.000	1.000	6.542	0.830
1	SampEn 0.10 50min	0.683	0.683	1.461	0.688
	Ppmd p5min	0.982	0.982	7.132	0.585
	Rácio LF /HF	0.997	0.997	1.002	0.698
	ApEn 0.10 p20min	0.843	0.843	1.118	0.695
2	SampEn 0.10 50min	0.679	0.679	1.584	0.556
	Rácio LF /HF	0.990	0.976	0.498	0.576
	ApEn 0.10 p20min	0.676	0.676	0.018	0.585

O valor próprio desta função discriminante é 0.708 e corresponde a 100% da variância explicada no que diz respeito às diferenças entre os grupos (Tabela 37). O valor da correlação canônica entre a função discriminante e os grupos é de 0.644, o que indica que o *effect size* do modelo encontrado é 0.41 (0.644^2).

Tabela 37 - Valor próprio. Fonte: elaboração própria.

Função	Valor próprio	% de variância	% cumulativa	Correlação canônica
1	0.708	100	100.0	0.644

O teste Lambda de Wilks representado na Tabela 38 indica que a função discriminante obtida consegue diferenciar os grupos de forma estatisticamente significativa ($p\text{-value} < 0.001$) o que indica que estamos perante um modelo forte.

Tabela 38 - Lambda Wilks. Fonte: elaboração própria.

Teste de funções	Lambda de Wilks	Qui-quadrado	df	p-value
1	0.585	16.599	2	<0.001

Na Tabela 39 encontra-se a variável com maior peso na função discriminante, neste caso, a Low Frequency (0.833). A variável com maior correlação com a função discriminante canônica padronizada é também a Low Frequency (0.741).

Tabela 39 - Coeficientes padronizados da função discriminantes e matriz de estrutura. Fonte: elaboração própria.

Variáveis	Coeficientes	Matriz de estrutura
SampEn 0.10 50min ^a	-	0.374
Low Frequency	0.833	0.741
Ppmd p5min	-0.678	-0.565
Rácio LF /HF ^a	-	0.096
ApEn 0.10 p20min ^a	-	0.568
a. Variável não utilizada na análise.		

Na Tabela 40 são apresentados os valores das funções de Fisher não padronizadas e é através deste valor que se escreve a função discriminante.

Tabela 40 - Coeficientes da função de Fisher. Fonte: elaboração própria.

Variáveis	Função 1
Low Frequency	0.160
Ppmd p5min	-0.309
(Constante)	1.973
Coeficientes não padronizados	

Assim, a função discriminante de Fisher para o quarto modelo é dada por:

$$Z = 1.973 + 0.160*(\text{Low Frequency}) - 0.309*(\text{Ppmd p5min})$$

Na Tabela 41 encontram-se os coeficientes das funções score discriminante para cada grupo, que permitem classificar os novos elementos.

Tabela 41 - Coeficientes das funções de classificação. Fonte: elaboração própria.

	Não académicos	Académicos
Low Frequency	0.340	0.079
Ppmd p5min	1.995	2.501
(Constante)	-12.864	-16.086
Funções discriminantes lineares de Fisher		

Funções score discriminante de Fisher:

$$Z_{\text{não académicos}} = -12.864 + 0.340*(\text{Low Frequency}) + 1.995*(\text{Ppmd p5min})$$

$$Z_{\text{acadêmicos}} = -16.086 + 0.079 * (\text{Low Frequency}) + 2.501 * (\text{Ppmd p5min})$$

A Tabela 42 apresenta os resultados da classificação usando o método *Stepwise*. Neste caso, o modelo conseguiu classificar corretamente 79.4% dos casos, o que representa uma classificação razoável. É ainda possível verificar a performance do modelo pela técnica de validação cruzada *leave-one-out*, que resultou em 25 (73.5%) casos corretamente classificados.

Tabela 42 - Resultados de classificação original e com validação cruzada. Fonte: elaboração própria.

			Associação ao grupo previsto		Total
			Não acadêmicos	Acadêmicos	
Original ^a	Contagem	Não acadêmicos	13	4	17
		Acadêmicos	3	14	17
	%	Não acadêmicos	76.5	23.5	100
		Acadêmicos	17.6	82.4	100
Com validação cruzada ^b	Contagem	Não acadêmicos	13	4	17
		Acadêmicos	5	12	17
	%	Não acadêmicos	76.5	23.5	100
		Acadêmicos	29.4	70.6	100

a. 79.4% de casos agrupados originais foram bem classificados.

b. 73.5% dos casos com validação cruzada foram bem classificados.

3.2. Classificação do conjunto de teste

Após encontradas as funções que classificam novos elementos, calculam-se as classes previstas para o conjunto de teste composto por 6 casos cujos valores nas variáveis do modelo se apresentam na Tabela 43.

Tabela 43 - Variáveis para os elementos do conjunto de teste. Fonte: elaboração própria.

		Casos acadêmicos			Casos não acadêmicos		
		Caso 8	Caso 176	Caso 202	Caso 44	Caso 144	Caso 177
Primeiro conjunto de variáveis	Variáveis						
	Gzip p5min	14.999	12.063	13.728	11.867	10.500	17.593
	SampEn 0.10 p20min	0.170	0.893	0.413	0.862	0.297	0.367
	ApEn 0.10 50min	0.170	0.544	0.614	0.647	0.302	0.548
	Low Frequency	0.782	8.437	5.315	10.380	1.702	6.614
	Rácio LF /HF	3.182	2.464	2.706	3.408	4.658	3.623
conjunto de	SampEn 0.10 50min	0.118	0.276	0.384	0.423	0.193	0.360

Low Frequency	0.782	8.437	5.315	10.380	1.702	6.614
Ppmd p5min	13.669	9.063	11.952	9.927	8.563	15.298
Rácio LF /HF	3.182	2.464	2.706	3.408	4.658	3.623
ApEn 0.10 p20min	0.255	1.072	0.638	1.069	0.479	0.546

Os resultados dos scores discriminantes para cada elemento do conjunto de teste são apresentados na Tabela 44, com realce para o score máximo obtido em cada caso e em cada modelo construído.

Tabela 44 - Score das funções discriminantes para cada caso do conjunto de teste pelos dois métodos de AD: *variáveis independentes juntas* versus *Stepwise*. Fonte: elaboração própria.

	Primeiro conjunto de variáveis				Segundo conjunto de variáveis			
	<i>Independentes juntas</i>		<i>Stepwise</i>		<i>Independentes juntas</i>		<i>Stepwise</i>	
	Não acadêmico	Acadêmico	Não acadêmico	Acadêmico	Não acadêmico	Acadêmico	Não acadêmico	Acadêmico
Caso 8 (acadêmico)	21.7499	25.7249	14.9811	18.8337	27.012	31.005	14.672	18.162
Caso 176 (acadêmico)	24.1271	24.4858	12.6082	12.2862	30.288	30.694	8.085	7.25
Caso 202 (acadêmico)	22.4302	22.8468	16.5680	16.6274	29.139	30.693	12.787	14.226
Caso 44 (não acadêmico)	25.0987	24.1460	13.1712	12.0226	34.143	33.806	10.470	9.561
Caso 144 (não acadêmico)	15.6126	16.1408	7.2655	7.8211	18.274	18.799	4.798	5.465
Caso 177 (não acadêmico)	32.9467	35.3179	23.6258	26.1921	39.439	42.457	19.904	22.697

A partir da Tabela 44, é possível verificar que o método de *variáveis independentes juntas* classificou corretamente os mesmos quatro dos seis casos tanto no primeiro conjunto de variáveis como no segundo. Já o modelo *Stepwise* classificou erradamente os mesmos três dos seis casos em ambos os conjuntos.

4. Discussão

Nos dois conjuntos de variáveis selecionados para a análise discriminante e nos dois métodos aplicados (*variáveis independentes juntas* e *Stepwise*) ao conjunto de treino, obtiveram-se boas percentagens de classificação dos casos nos grupos académicos ou não académicos

No primeiro conjunto de variáveis (Gzip p5min, SampEn 0.10 p20min, ApEn 0.10 50min, LF e o Rácio LF/HF), o método das *variáveis independentes juntas* classificou 85.3% dos casos corretamente e através da técnica de validação cruzada *leave-one-out*, a percentagem de acerto foi de 79.4%. O método *Stepwise* classificou 85.3% corretamente e obteve exatamente a mesma percentagem através da técnica *leave-one-out*.

No segundo conjunto de variáveis (SampEn 0.10 50min, LF, Pmpd p5min, Rácio LF/HF e ApEn 0.10 p20min), o método das *variáveis independentes juntas* obteve 85.3% dos casos classificados corretamente, e na técnica *leave-one-out* essa percentagem foi de 79.4% (exatamente igual ao obtido no primeiro conjunto de variáveis). Com recurso ao método *Stepwise*, a AD classificou de forma correta 79.4% e, utilizando a *leave-one-out* a percentagem foi de 73.5%, ou seja, inferior aos valores obtidos com o primeiro conjunto de variáveis).

Quando se utilizou o conjunto de teste para verificar qual a classificação atribuída a um novo elemento, conclui-se que, para ambos os conjuntos de variáveis, o método das *variáveis independentes juntas* apresentou maior sucesso. Em seis casos novos, o método das *variáveis independentes juntas* classificou corretamente 4 e erradamente 2. Enquanto que o método *Stepwise* classificou corretamente 3 casos, o que resulta numa percentagem de concordância de 50%.

No entanto, é de salientar que o valor do Lambda de Wilks das dez variáveis que compõem os dois conjuntos em estudo está longe de 0, indicando por isso que o seu poder discriminante não é elevado, o que pode justificar classificações incorretas.

No que diz respeito ao teste de igualdade de matrizes de covariância entre os dois grupos, no primeiro conjunto de variáveis, foi rejeitada a igualdade ($p\text{-value} = 0.020$). Tendo em conta que este é um dos pressupostos da AD, este ponto representa uma limitação do modelo apresentado para a previsão de acidemia fetal. Já no segundo conjunto de variáveis, o teste de igualdade de matrizes de covariância entre os grupos permitiu não rejeitar que existe homogeneidade das matrizes de covariância ($p\text{-value} = 0.065$).

No futuro, dispondo de um conjunto de dados de maior dimensão e mantendo a homogeneidade de tamanhos das amostras, o pressuposto da igualdade das matrizes de covariância poderá ser cumprido.

Apesar de alguns pressupostos não se verificarem, as percentagens de classificações corretas podem sempre servir como guia, para aceitar como bom, qualquer modelo de descrição. Sendo esta percentagem elevada, a violação dos pressupostos não é grave. No caso dessa percentagem ser pequena não existiria evidência suficiente para distinguir se tal ocorreu devido ao facto de não se terem cumprido os pressupostos da AD, ou se o conjunto de dados teria fraco poder para discriminar os indivíduos.

O SPSS® mostrou-se, ao longo do trabalho, uma ferramenta muito poderosa no que diz respeito aos recursos que disponibiliza para realizar a AD.

5. Conclusão e trabalho futuro

A importância deste estudo prende-se com a criação de um modelo que permita a deteção antecipada de fetos com acidemia. A hipoxia fetal pode conduzir a sérias consequências para o recém-nascido, nomeadamente sequelas cerebrais ou mesmo fatais. A intervenção precoce, que depende do diagnóstico, é de extrema importância na prevenção de sequelas no feto. Algumas dificuldades ocorreram ao longo deste trabalho, sendo que a principal dificuldade está no número reduzido de casos académicos disponível na amostra. Por um lado, limitou a possibilidade de escolha na técnica a utilizar para abordar o problema, por outro lado, obrigou a definir um conjunto de teste de tamanho reduzido, dificultando a avaliação da capacidade de generalização do modelo.

No teste de igualdade de matrizes de covariância entre os grupos para o primeiro conjunto de variáveis analisadas, foi rejeitada a igualdade, o que representa uma limitação do modelo apresentado para a previsão de acidemia fetal.

Para continuação do trabalho realizado deveria-se, primeiramente, verificar se existe mais algum conjunto de variáveis que não estejam correlacionadas e efetuar a AD para perceber se mais variáveis classificam de igual ou melhor modo os casos e novos eventos.

Se esta metodologia fosse testada num outro conjunto de dados, de maior dimensão e, mantendo a homogeneidade dos tamanhos das amostras, o pressuposto da igualdade das matrizes de covariância poderia ser cumprido. Era também interessante aplicar diretamente os resultados obtidos em outras bases de dados.

Teria certamente interesse utilizar adicionalmente ferramentas de *Machine Learning*, nomeadamente Árvores de Decisão, método esse muito utilizado em tarefas de classificação e regressão.

Referências

1. Pereira, S.N., *Qualidade em Saúde: o Papel dos Sistemas de Informação* in *Departamento de Física*. 2008/2009, Faculdade de Ciências e Tecnologia da Universidade de Lisboa.
2. Task Force on Neonatal Encephalopathy and Cerebral Palsy. *Neonatal encephalopathy and cerebral palsy: defining the pathogenesis and pathophysiology*. 2003. Washington, DC.
3. Practice, A.C.O.O., *ACOG Committee Opinion No. 348, November 2006: Umbilical cord blood gas and acid-base analysis*. *Obstet Gynecol*, 2006. **108**(5): p. 1319-22.
4. *Confidential Enquiry Into Maternal and Child Health. Perinatal Mortality Surveillance, 2004: England, Wales and Northern Ireland*. 2006, London: CEMACH.
5. Pomerance, J.J., *Interpreting Umbilical Cord Blood Gases: For Clinicians Caring for the Fetus Or Newborn*. 2004: Pasadena, Calif. : BNMG.
6. Bobrow, C.S. and P.W. Soothill, *Causes and consequences of fetal acidosis*. *Arch Dis Child Fetal Neonatal Ed*, 1999. **80**(3): p. F246-9.
7. Sousa, P., *Omniview-SisPorto®: Análise computadorizada dos sinais de monitorizar fetal Interface STAN®*. 2011, Faculdade de Engenharia da Universidade do Porto.
8. CESDI, *4th annual report*. 1997: London, UK: Maternal and Child Health Research Consortium.
9. Ayres-de-Campos, D., et al., *Inconsistencies in classification by experts of cardiotocograms and subsequent clinical decision*. *Br J Obstet Gynaecol*, 1999. **106**(12): p. 1307-10.
10. Bernardes, J., et al., *Evaluation of interobserver agreement of cardiotocograms*. *Int J Gynaecol Obstet*, 1997. **57**(1): p. 33-7.
11. Ayres de Campos, D., *Desenvolvimento, Avaliação Clínica e Optimização de um Sistema de Análise Automatizada do Cardiotocograma*. 2001.
12. Lotgering, F.K., H.C. Wallenburg, and H.J. Schouten, *Interobserver and intraobserver variation in the assessment of antepartum cardiotocograms*. *Am J Obstet Gynecol*, 1982. **144**(6): p. 701-5.
13. Searle, J.R., et al., *Computerized analysis of resting fetal heart rate tracings*. *Obstet Gynecol*, 1988. **71**(3 Pt 1): p. 407-11.
14. Dawes, G.S., C.W. Redman, and J.H. Smith, *Improvements in the registration and analysis of fetal heart rate records at the bedside*. *Br J Obstet Gynaecol*, 1985. **92**(4): p. 317-25.
15. Ayres-de-Campos, D., et al., *Prediction of neonatal state by computer analysis of fetal heart rate tracings: the antepartum arm of the SisPorto multicentre validation study*. *Eur J Obstet Gynecol Reprod Biol*, 2005. **118**(1): p. 52-60.
16. Reis, E., *Estatística Multivariada Aplicada*. Edições Sílabo ed. 1997, Lisboa.
17. Fisher, R.A., *The use of multiple measurements in taxonomic problems*. *Annals of Eugenics*, 1936. **7**: p. 179-188.
18. Marôco, J., *Análise estatística : com utilização do SPSS*. 3 ed. 2007.
19. Hair, J.F., et al., *Análise multivariada de dados*. 2005, Porto Alegre: Bookman.
20. Sharma, S., *Applied Multivariate Techniques*. 1996, New York: J. Wiley & Sons.
21. Fávero, L.P., et al., *Análise de dados: modelagem multivariada para tomada de decisões*. 2009, Rio de Janeiro: Elsevier.
22. Gonçalves, H., et al., *Internal versus external intrapartum foetal heart rate monitoring: the effect on linear and nonlinear parameters*. *Physiol Meas*, 2006. **27**(3): p. 307-19.
23. Gonçalves, H., et al., *Linear and nonlinear fetal heart rate analysis of normal and acidemic fetuses in the minutes preceding delivery*. *Med Biol Eng Comput*, 2006. **44**(10): p. 847-55.
24. Chua, K.C., et al., *Automatic identification of epilepsy by HOS and power spectrum parameters using EEG signals: a comparative study*. *Conf Proc IEEE Eng Med Biol Soc*, 2008. **2008**: p. 3824-7.

25. Signorini, M.G., et al., *Linear and nonlinear parameters for the analysis of fetal heart rate signal from cardiotocographic recordings*. IEEE Trans Biomed Eng, 2003. **50**(3): p. 365-74.
26. Chung, D.Y., et al., *Spectral analysis of fetal heart rate variability as a predictor of intrapartum fetal distress*. Int J Gynaecol Obstet, 2001. **73**(2): p. 109-16.
27. Ayres-de Campos, D., et al., *SisPorto 2.0: a program for automated analysis of cardiotocograms*. J Matern Fetal Med, 2000. **9**(5): p. 311-8.
28. Ayres-de-Campos, D., et al., *Omniview-SisPorto 3.5 - a central fetal monitoring station with online alerts based on computerized cardiotocogram+ST event analysis*. J Perinat Med, 2008. **36**(3): p. 260-4.
29. Bernardes, J., et al., *The Porto system for automated cardiotocographic signal analysis*. J Perinat Med, 1991. **19**(1-2): p. 61-5.
30. Pincus, S.M., I.M. Gladstone, and R.A. Ehrenkranz, *A regularity statistic for medical data analysis*. J Clin Monit, 1991. **7**(4): p. 335-45.
31. Richman, J.S. and J.R. Moorman, *Physiological time-series analysis using approximate entropy and sample entropy*. Am J Physiol Heart Circ Physiol, 2000. **278**(6): p. H2039-49.
32. Pincus, S.M., *Approximate entropy as a measure of system complexity*. Proc Natl Acad Sci U S A, 1991. **88**(6): p. 2297-301.
33. Chen, X., I. Solomon, and K. Chon, *Comparison of the use of approximate entropy and sample entropy: applications to neural respiratory signal*. Conf Proc IEEE Eng Med Biol Soc, 2005. **4**: p. 4212-5.
34. Jiang, G.J., et al., *Sample entropy analysis of EEG signals via artificial neural networks to model patients' consciousness level based on anesthesiologists experience*. Biomed Res Int, 2015. **2015**: p. 343478.
35. Fusheng, Y., et al., *Approximate entropy and its application in biosignal analysis*. Nonlinear Biomedical Signal Processing: Dynamic Analysis and Modeling, Volume 2. 2012, New York: Metin Akay.
36. Pincus, S., *Approximate entropy (ApEn) as a complexity measure*. Chaos, 1995. **5**(1): p. 110-117.
37. Santos, C.a.B., J. and Vitanyi, P. M. B. and Antunes, L., *Clustering fetal heart rate tracings by compression*. 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06), 2006. **2006**: p. 685-690.
38. Costa, A., et al., *Prediction of neonatal acidemia by computer analysis of fetal heart rate and ST event signals*. Am J Obstet Gynecol, 2009. **201**(5): p. 464.e1-6.
39. Brillinger, D.R., *Time Series: Data Analysis and Theory*. 2nd ed. 1981, San Francisco.

Anexos

Tabela 45 - Correlações de Spearman entre as variáveis consideradas significativas após o teste de Mann-Whitney. Fonte: elaboração própria.

		ApEn n 0.10 50m in	Low Frequ ency	Sam pEn 0.10 50m in	Ver y Low Frequ ency	ApE n 0.20 50m in	Ppm d p5m in	ApE n 0.10 u20 min	ApE n 0.15 50m in	Bzip p5m in	Sam pEn 0.15 50m in	ApE n 0.15 u20 min	Sam pEn 0.15 u20 min	Sam pEn 0.10 u20 min	ApE n 0.10 u10 min	Gzip p5m in	Sam pEn 0.20 50m in	ApE n 0.10 p20 min	Sam pEn 0.20 u20 min	ApE n 0.10 p10 min	ApE n 0.10 u5m in	Lzm a p5m in	Paq 8l p5m in	ApE n 0.15 p20 min	Sam pEn 0.10 p20 min	Gzip p10 min	ApE n 0.15 p5m in	ApE n 0.15 u10 min	Bzip p10 min	Hig h freq uency	Raci o LF /HF	ApE n 0.20 u20 min	Ppm d p10 min
ApEn 0.10 50min	r	1.000	.750 ⁺	.970 ⁺	.750 ⁺⁺	.888 ⁺	- 0.074	.809 ⁺	.836 ⁺	- 0.043	.858 ⁺	.870 ⁺	.875 ⁺	.783 ⁺	.728 ⁺	0.001	.899 ⁺	.547 ⁺	.837 ⁺	.475 ⁺	.735 ⁺	0.011	0.038	.540 ⁺	.576 ⁺	- 0.042	.486 ⁺	.785 ⁺	- 0.077	.697 ⁺	0.095	.820 ⁺	- 0.094
	Si g.		0.000	0.000	0.000	0.000	0.677	0.000	0.000	0.809	0.000	0.000	0.000	0.000	0.997	0.000	0.001	0.000	0.005	0.000	0.951	0.830	0.001	0.000	0.813	0.004	0.000	0.664	0.000	0.592	0.000	0.596	
Low Frequenc y	r	.750 ⁺⁺	1.000	.753 ⁺⁺	.994 ⁺⁺	.822 ⁺⁺	- 0.085	.620 ⁺⁺	.692 ⁺⁺	- 0.042	.669 ⁺⁺	.670 ⁺⁺	.694 ⁺⁺	.640 ⁺⁺	.617 ⁺⁺	- 0.047	.761 ⁺⁺	.498 ⁺⁺	.644 ⁺⁺	0.297 ⁺⁺	.575 ⁺⁺	- 0.059	0.025	.493 ⁺⁺	.566 ⁺⁺	- 0.069	.387 ⁺⁺	.540 ⁺⁺	- 0.080	.884 ⁺⁺	0.177	.651 ⁺⁺	- 0.094
	Si g.	0.000		0.000	0.000	0.000	0.633	0.000	0.000	0.812	0.000	0.000	0.000	0.000	0.792	0.000	0.003	0.000	0.088	0.000	0.742	0.889	0.003	0.000	0.697	0.024	0.001	0.653	0.000	0.317	0.000	0.597	
SamEn 0.10 50min	r	.970 ⁺⁺	.753 ⁺⁺	1.000	.748 ⁺⁺	.881 ⁺⁺	- 0.050	.792 ⁺⁺	.819 ⁺⁺	- 0.018	.884 ⁺⁺	.823 ⁺⁺	.856 ⁺⁺	.814 ⁺⁺	.701 ⁺⁺	0.020	.922 ⁺⁺	.534 ⁺⁺	.806 ⁺⁺	.472 ⁺⁺	.702 ⁺⁺	0.030	0.059	.517 ⁺⁺	.594 ⁺⁺	- 0.013	.508 ⁺⁺	.729 ⁺⁺	- 0.035	.692 ⁺⁺	0.096	.767 ⁺⁺	- 0.054
	Si g.	0.000	0.000		0.000	0.000	0.779	0.000	0.000	0.917	0.000	0.000	0.000	0.000	0.911	0.000	0.001	0.000	0.005	0.000	0.864	0.738	0.002	0.000	0.942	0.002	0.000	0.843	0.000	0.589	0.000	0.763	
Very Low Frequenc y	r	.750 ⁺⁺	.994 ⁺⁺	.748 ⁺⁺	1.000	.822 ⁺⁺	0.071	.627 ⁺⁺	.700 ⁺⁺	0.024	.663 ⁺⁺	.673 ⁺⁺	.689 ⁺⁺	.644 ⁺⁺	.628 ⁺⁺	0.043	.756 ⁺⁺	.488 ⁺⁺	.653 ⁺⁺	0.303	.583 ⁺⁺	0.058	0.011	.509 ⁺⁺	.554 ⁺⁺	0.074	.385 ⁺⁺	.550 ⁺⁺	0.071	.921 ⁺⁺	0.094	.668 ⁺⁺	0.092
	Si g.	0.000	0.000	0.000		0.000	0.688	0.000	0.000	0.892	0.000	0.000	0.000	0.000	0.811	0.000	0.003	0.000	0.082	0.000	0.745	0.949	0.002	0.001	0.677	0.025	0.001	0.688	0.000	0.596	0.000	0.606	
ApEn 0.20 50min	r	.888 ⁺⁺	.822 ⁺⁺	.881 ⁺⁺	.822 ⁺⁺	1.000	0.053	.815 ⁺⁺	.848 ⁺⁺	0.084	.825 ⁺⁺	.870 ⁺⁺	.893 ⁺⁺	.808 ⁺⁺	.752 ⁺⁺	0.127	.958 ⁺⁺	.522 ⁺⁺	.826 ⁺⁺	.431 ⁺⁺	.724 ⁺⁺	0.136	0.151	.512 ⁺⁺	.573 ⁺⁺	0.113	.448 ⁺⁺	.751 ⁺⁺	0.065	.780 ⁺⁺	0.038	.819 ⁺⁺	0.043
	Si g.	0.000	0.000	0.000	0.000		0.767	0.000	0.000	0.638	0.000	0.000	0.000	0.000	0.473	0.000	0.002	0.000	0.011	0.000	0.444	0.396	0.002	0.000	0.526	0.008	0.000	0.716	0.000	0.831	0.000	0.808	
Ppm d p5min	r	- 0.074	- 0.085	- 0.050	0.071	0.053	1.000	- 0.055	0.015	.989 ⁺⁺	- 0.015	0.014	0.038	0.000	0.164	.947 ⁺⁺	0.034	.441 ⁺⁺	0.101	.464 ⁺⁺	0.163	.917 ⁺⁺	.959 ⁺⁺	- 0.212	- 0.356	.873 ⁺⁺	- 0.435	0.104	.921 ⁺⁺	0.062	- 0.147	0.082	.903 ⁺⁺
	Si g.	0.677	0.633	0.779	0.688	0.767		0.759	0.935	0.000	0.934	0.939	0.830	0.998	0.353	0.000	0.848	0.009	0.569	0.006	0.358	0.000	0.000	0.229	0.039	0.000	0.010	0.557	0.000	0.729	0.406	0.645	0.000
ApEn 0.10 u20min	r	.809 ⁺⁺	.620 ⁺⁺	.792 ⁺⁺	.627 ⁺⁺	.815 ⁺⁺	- 0.055	1.000	.865 ⁺⁺	- 0.032	.846 ⁺⁺	.937 ⁺⁺	.929 ⁺⁺	.948 ⁺⁺	.845 ⁺⁺	- 0.016	.829 ⁺⁺	.433 ⁺⁺	.924 ⁺⁺	.423 ⁺⁺	.761 ⁺⁺	- 0.024	0.029	.547 ⁺⁺	.460 ⁺⁺	- 0.045	0.302	.864 ⁺⁺	0.055	.599 ⁺⁺	0.030	.898 ⁺⁺	- 0.091
	Si g.	0.000	0.000	0.000	0.000	0.000	0.759		0.000	0.856	0.000	0.000	0.000	0.000	0.930	0.000	0.010	0.000	0.013	0.000	0.893	0.871	0.001	0.006	0.799	0.083	0.000	0.759	0.000	0.864	0.000	0.609	
ApEn 0.15 50min	r	.836 ⁺⁺	.692 ⁺⁺	.819 ⁺⁺	.700 ⁺⁺	.848 ⁺⁺	0.015	.865 ⁺⁺	1.000	0.037	.960 ⁺⁺	.878 ⁺⁺	.881 ⁺⁺	.855 ⁺⁺	.745 ⁺⁺	0.006	.867 ⁺⁺	.449 ⁺⁺	.916 ⁺⁺	.345 ⁺⁺	.691 ⁺⁺	0.010	0.066	.592 ⁺⁺	.484 ⁺⁺	- 0.029	0.236	.750 ⁺⁺	0.043	.656 ⁺⁺	0.104	.897 ⁺⁺	- 0.077
	Si g.	0.000	0.000	0.000	0.000	0.000	0.935	0.000		0.838	0.000	0.000	0.000	0.000	0.975	0.000	0.008	0.000	0.046	0.000	0.956	0.713	0.000	0.004	0.873	0.179	0.000	0.809	0.000	0.559	0.000	0.667	
Bzip p5min	r	0.043	0.042	0.018	0.024	0.084	.989 ⁺⁺	0.032	0.037	1.000	0.005	- 0.001	0.052	0.026	0.140	.935 ⁺⁺	0.065	.440 ⁺⁺	0.115	.456 ⁺⁺	0.151	.910 ⁺⁺	.967 ⁺⁺	0.197	.345 ⁺⁺	.841 ⁺⁺	- 0.423	0.089	.910 ⁺⁺	0.125	- 0.201	0.107	.878 ⁺⁺
	Si g.	0.809	0.812	0.917	0.892	0.638	0.000	0.856	0.838		0.977	0.995	0.768	0.882	0.429	0.000	0.714	0.009	0.516	0.007	0.394	0.000	0.000	0.264	0.046	0.000	0.013	0.615	0.000	0.481	0.254	0.547	0.000
SamEn 0.15 u20min	r	.858 ⁺⁺	.669 ⁺⁺	.884 ⁺⁺	.663 ⁺⁺	.825 ⁺⁺	- 0.015	.846 ⁺⁺	.960 ⁺⁺	0.005	1.000	.844 ⁺⁺	.865 ⁺⁺	.874 ⁺⁺	.708 ⁺⁺	0.000	.896 ⁺⁺	.488 ⁺⁺	.878 ⁺⁺	.396 ⁺⁺	.677 ⁺⁺	0.011	0.054	.590 ⁺⁺	.541 ⁺⁺	- 0.033	0.332	.730 ⁺⁺	0.053	.598 ⁺⁺	0.147	.836 ⁺⁺	- 0.079
	Si g.	0.000	0.000	0.000	0.000	0.000	0.934	0.000	0.000	0.977		0.000	0.000	0.000	0.999	0.000	0.003	0.000	0.021	0.000	0.950	0.761	0.000	0.001	0.852	0.055	0.000	0.768	0.000	0.408	0.000	0.658	

ApEn 0.15 u20min	r	,870	,670	,823	,673 ⁺⁺	,870	-, 0.014	,937	,878	-, 0.001	,844	1.000	,971	,877	,843	0.025	,872	,448	,934	,386	,813	0.010	0.054	,498	,465	-, 0.022	0.281	,913	-, 0.069	,622	0.102	,942	-, 0.089
	Si g-	0.000	0.000	0.000	0.000	0.939	0.000	0.000	0.000	0.995	0.000		0.000	0.000	0.000	0.889	0.000	0.008	0.000	0.024	0.000	0.956	0.761	0.003	0.006	0.904	0.107	0.000	0.699	0.000	0.566	0.000	0.615
SampEn 0.15 u20min	r	,875	,694	,856	,689 ⁺⁺	,893	0.038	,929	,881	0.052	,865	,971	1.000	,912	,806	0.081	,906	,476	,937	,411	,761	0.067	0.108	,468	,504	0.032	0.285	,861	-, 0.012	,620	0.145	,906	-, 0.035
	Si g-	0.000	0.000	0.000	0.000	0.000	0.830	0.000	0.000	0.768	0.000	0.000		0.000	0.000	0.647	0.000	0.004	0.000	0.016	0.000	0.705	0.543	0.005	0.002	0.858	0.103	0.000	0.947	0.000	0.414	0.000	0.846
SampEn 0.10 u20min	r	,783	,640	,814	,644 ⁺⁺	,808	0.000	,948	,855	0.026	,874	,877	,912	1.000	,832	0.034	,862	,449	,887	,457	,725	0.018	0.077	,505	,510	-, 0.008	0.294	,809	-, 0.012	,607	0.026	,828	-, 0.046
	Si g-	0.000	0.000	0.000	0.000	0.000	0.998	0.000	0.000	0.882	0.000	0.000	0.000		0.000	0.847	0.000	0.008	0.000	0.007	0.000	0.917	0.664	0.002	0.002	0.965	0.092	0.000	0.948	0.000	0.886	0.000	0.795
ApEn 0.10 u10min	r	,728	,617	,701	,628 ⁺⁺	,752	-, 0.164	,845	,745	-, 0.140	,708	,843	,806	,832	1.000	-, 0.087	,740	,573	,798	,579	,923	-, 0.110	-, 0.082	-, ,632	,594	-, 0.125	,353	,918	-, 0.196	,586	0.041	,812	-, 0.207
	Si g-	0.000	0.000	0.000	0.000	0.000	0.353	0.000	0.000	0.429	0.000	0.000	0.000	0.000		0.624	0.000	0.000	0.000	0.000	0.000	0.536	0.646	0.000	0.000	0.483	0.041	0.000	0.266	0.000	0.819	0.000	0.241
Gzip p5min	r	0.001	-, 0.047	0.020	0.043	0.127	,947 ⁺⁺	-, 0.016	0.006	,935 ⁺⁺	0.000	0.025	0.081	0.034	-, 0.087	1.000	0.106	-, 0.326	0.121	-, 0.299	-, 0.056	,979 ⁺⁺	,975 ⁺⁺	-, 0.149	-, 0.238	,944 ⁺⁺	-, 0.240	-, 0.013	,922 ⁺⁺	0.084	-, 0.156	0.105	,925 ⁺⁺
	Si g-	0.997	0.792	0.911	0.811	0.473	0.000	0.930	0.975	0.000	0.999	0.889	0.647	0.847	0.624		0.552	0.060	0.494	0.085	0.754	0.000	0.000	0.400	0.175	0.000	0.172	0.943	0.000	0.638	0.378	0.556	0.000
SampEn 0.20 50min	r	,899	,761	,922	-, ,756 ⁺⁺	,958	0.034	,829	,867	0.065	,896	,872	,906	,862	,740	0.106	1.000	,527	,832	,454	,730	0.112	0.128	,492	,598	0.077	,456	,756	0.027	,703	0.056	,807	0.008
	Si g-	0.000	0.000	0.000	0.000	0.000	0.848	0.000	0.000	0.714	0.000	0.000	0.000	0.000	0.000	0.552		0.001	0.000	0.007	0.000	0.528	0.472	0.003	0.000	0.666	0.007	0.000	0.879	0.000	0.751	0.000	0.965
ApEn 0.10 p20min	r	,547	,498	,534	-, ,488 ⁺⁺	,522	-, ,441 ⁺⁺	,433	,449	-, ,440 ⁺⁺	,488	,448	,476	,449	,573	-, 0.326	,527	1.000	,446	,850	,594	-, 0.333	-, ,361	,773	,969	-, ,396	,537	,521	-, ,479 ⁺⁺	,370	0.215	,376	-, ,474 ⁺⁺
	Si g-	0.001	0.003	0.001	0.003	0.002	0.009	0.010	0.008	0.009	0.003	0.008	0.004	0.008	0.000	0.060	0.001		0.008	0.000	0.000	0.054	0.036	0.000	0.000	0.020	0.001	0.002	0.004	0.031	0.222	0.028	0.005
SampEn 0.20 u20min	r	,837	,644	,806	-, ,653 ⁺⁺	,826	0.101	,924	,916	0.115	,878	,934	,937	,887	,798	0.121	,832	,446	1.000	,387	,763	0.107	0.161	,554	,479	0.038	0.194	,862	0.016	,625	0.081	,964	-, 0.021
	Si g-	0.000	0.000	0.000	0.000	0.000	0.569	0.000	0.000	0.516	0.000	0.000	0.000	0.000	0.000	0.494	0.000	0.008		0.024	0.000	0.545	0.362	0.001	0.004	0.830	0.272	0.000	0.929	0.000	0.651	0.000	0.908
ApEn 0.10 p10min	r	,475	0.297	,472	-, 0.303	,431	-, ,464 ⁺⁺	,423	,345	-, ,456 ⁺⁺	,396	,386	,411	,457	,579	-, 0.299	,454	,850	,387	1.000	,564	-, 0.307	-, ,340	,655	,812	-, ,352	,620	,552	-, ,459 ⁺⁺	0.234	0.043	0.324	-, ,460 ⁺⁺
	Si g-	0.005	0.088	0.005	0.082	0.011	0.006	0.013	0.046	0.007	0.021	0.024	0.016	0.007	0.000	0.085	0.007	0.000	0.024		0.001	0.078	0.049	0.000	0.000	0.041	0.000	0.001	0.006	0.184	0.811	0.062	0.006
ApEn 0.10 u5min	r	,735	,575	,702	-, ,583 ⁺⁺	,724	-, 0.163	,761	,691	-, 0.151	,677	,813	,761	,725	,923	-, 0.056	,730	,594	,763	,564	1.000	-, 0.065	-, 0.067	,631	,607	-, 0.118	,401	,916	-, 0.208	,531	0.052	,778	-, 0.206
	Si g-	0.000	0.000	0.000	0.000	0.000	0.358	0.000	0.000	0.394	0.000	0.000	0.000	0.000	0.000	0.754	0.000	0.000	0.000	0.001		0.714	0.708	0.000	0.000	0.505	0.019	0.000	0.238	0.001	0.768	0.000	0.243
Lzma p5min	r	0.011	-, 0.059	0.030	0.058	0.136	,917 ⁺⁺	-, 0.024	0.010	,910	0.011	0.010	0.067	0.018	-, 0.110	,979	0.112	-, 0.333	0.107	-, 0.307	-, 0.065	1.000	,970	-, 0.160	-, 0.253	,931	-, 0.191	-, 0.025	,912	0.067	-, 0.150	0.093	,914
	Si g-	0.951	0.742	0.864	0.745	0.444	0.000	0.893	0.956	0.000	0.950	0.956	0.705	0.917	0.536	0.000	0.528	0.054	0.545	0.078	0.714		0.000	0.367	0.148	0.000	0.280	0.887	0.000	0.705	0.397	0.601	0.000
Paq8l p5min	r	0.038	-, 0.025	0.059	0.011	0.151	,959	0.029	0.066	,967	0.054	0.054	0.108	0.077	-, 0.082	,975	0.128	-, ,361	0.161	-, ,340	-, 0.067	,970	1.000	-, 0.121	-, 0.268	,894	-, 0.264	-, 0.005	,930	0.136	-, 0.206	0.153	,906
	Si g-	0.830	0.889	0.738	0.949	0.396	0.000	0.871	0.713	0.000	0.761	0.761	0.543	0.664	0.646	0.000	0.472	0.036	0.362	0.049	0.708	0.000		0.495	0.126	0.000	0.132	0.976	0.000	0.442	0.243	0.386	0.000

ApEn 0.15 p20min	r	.540	.493	.517 ⁺	.509 ⁺⁺	.512	.212	.547	.592	.197	.590	.498	.468	.505	.632	.149	.492	.773	.554	.655	.631	.160	.121	1.000	.788	.224	.422	.566	.226	.459	0.093	.524	.261
	Si g _e	0.001	0.003	0.002	0.002	0.002	0.229	0.001	0.000	0.264	0.003	0.005	0.002	0.000	0.400	0.003	0.000	0.001	0.000	0.000	0.367	0.495		0.000	0.204	0.013	0.000	0.198	0.006	0.602	0.001	0.136	
SumpEn 0.10 p20min	r	.576 ⁺	.566 ⁺	.594 ⁺	.554 ⁺⁺	.573 ⁺	.356 ⁺	.460 ⁺	.484 ⁺	.345 ⁺	.541 ⁺	.465 ⁺	.504 ⁺	.510 ⁺	.594 ⁺	.238	.598 ⁺	.969 ⁺	.479 ⁺	.812 ⁺	.607 ⁺	.253	.268	.788 ⁺	1.000	.313	.567 ⁺	.527 ⁺	.378 ⁺	.442 ⁺	0.183	.402	.385 ⁺
	Si g _e	0.000	0.000	0.000	0.001	0.000	0.039	0.006	0.004	0.046	0.001	0.006	0.002	0.002	0.000	0.175	0.000	0.000	0.004	0.000	0.000	0.148	0.126	0.000		0.071	0.000	0.001	0.027	0.009	0.299	0.018	0.025
Gzip p10min	r	0.042	0.069	0.013	0.074	0.113	.873 ⁺	.045	0.029	.841 ⁺	0.033	0.022	.032	0.008	0.125	.944 ⁺	0.077	.396 ⁺	0.038	.352 ⁺	0.118	.931 ⁺	.894 ⁺	0.224	0.313	1.000	.0139	0.090	.944 ⁺	0.025	0.095	0.030	.971 ⁺
	Si g _e	0.813	0.697	0.942	0.677	0.526	0.000	0.799	0.873	0.000	0.852	0.904	0.858	0.965	0.483	0.000	0.666	0.020	0.830	0.041	0.505	0.000	0.000	0.204	0.071		0.432	0.612	0.000	0.890	0.591	0.866	0.000
ApEn 0.15 p5min	r	.486 ⁺	.387 ⁺	.508 ⁺	.385 ⁺	.448 ⁺	.435 ⁺	0.302	0.236	.423 ⁺	0.332	0.281	0.285	0.294	.353 ⁺	.240	.456 ⁺	.537 ⁺	0.194	.620 ⁺	.401 ⁺	.0191	0.264	.422	.567 ⁺	.0139	1.000	.372	.210	0.327	.071	0.194	.203
	Si g _e	0.004	0.024	0.002	0.025	0.008	0.010	0.083	0.179	0.013	0.055	0.107	0.103	0.092	0.041	0.172	0.007	0.001	0.272	0.000	0.019	0.280	0.132	0.013	0.000	0.432		0.030	0.233	0.059	0.690	0.273	0.250
ApEn 0.15 u10min	r	.785 ⁺⁺	.540 ⁺⁺	.729 ⁺⁺	.550 ⁺⁺	.751 ⁺⁺	.0104	.864 ⁺⁺	.750 ⁺⁺	.089	.730 ⁺⁺	.913 ⁺⁺	.861 ⁺⁺	.809 ⁺⁺	.918 ⁺⁺	.013	.756 ⁺⁺	.521 ⁺⁺	.862 ⁺⁺	.552 ⁺⁺	.916 ⁺⁺	.025	.005	.566 ⁺⁺	.527 ⁺⁺	.090	.372 ⁺⁺	1.000	.0171	.523 ⁺⁺	.010	.885 ⁺⁺	.187
	Si g _e	0.000	0.001	0.000	0.001	0.000	0.557	0.000	0.000	0.615	0.000	0.000	0.000	0.000	0.000	0.943	0.000	0.002	0.000	0.001	0.000	0.887	0.976	0.000	0.001	0.612	0.030		0.335	0.002	0.956	0.000	0.291
Bzip p10min	r	0.077	0.080	0.035	0.071	0.065	.921 ⁺⁺	.055	.043	.910 ⁺⁺	.053	.069	.012	.012	.0196	.922 ⁺⁺	.027	.479 ⁺⁺	.016	.459 ⁺⁺	.208	.912 ⁺⁺	.930 ⁺⁺	.226	.378 ⁺⁺	.944 ⁺⁺	.0210	.0171	1.000	.063	.0187	.001	.986
	Si g _e	0.664	0.653	0.843	0.688	0.716	0.000	0.759	0.809	0.000	0.768	0.699	0.947	0.948	0.266	0.000	0.879	0.004	0.929	0.006	0.238	0.000	0.000	0.198	0.027	0.000	0.233	0.335		0.722	0.290	0.997	0.000
High freque ncy	r	.697 ⁺⁺	.884 ⁺⁺	.692 ⁺⁺	.921 ⁺⁺	.780 ⁺⁺	0.062	.599 ⁺⁺	.656 ⁺⁺	0.125	.598 ⁺⁺	.622 ⁺⁺	.620 ⁺⁺	.607 ⁺⁺	.586 ⁺⁺	0.084	.703 ⁺⁺	.370 ⁺⁺	.625 ⁺⁺	0.234	.531 ⁺⁺	0.067	0.136	.459 ⁺⁺	.442 ⁺⁺	0.025	0.327	.523 ⁺⁺	0.063	1.000	.0271	.679 ⁺⁺	0.024
	Si g _e	0.000	0.000	0.000	0.000	0.000	0.729	0.000	0.000	0.481	0.000	0.000	0.000	0.000	0.000	0.638	0.000	0.031	0.000	0.184	0.001	0.705	0.442	0.006	0.009	0.890	0.059	0.002	0.722		0.121	0.000	0.894
Ratio LF /HF	r	0.095	0.177	0.096	0.094	0.038	.147	0.030	0.104	.201	0.147	0.102	0.145	0.026	0.041	.156	0.056	0.215	0.081	0.043	0.052	.0150	.0206	0.093	0.183	.0095	.071	0.010	.0187	.0271	1.000	.014	.0126
	Si g _e	0.592	0.317	0.589	0.596	0.831	0.406	0.864	0.559	0.254	0.408	0.566	0.414	0.886	0.819	0.378	0.751	0.222	0.651	0.811	0.768	0.397	0.243	0.602	0.299	0.591	0.690	0.956	0.290	0.121		0.938	0.476
ApEn 0.20 u20min	r	.820	.651 ⁺	.767 ⁺	.668 ⁺⁺	.819	0.082	.898 ⁺	.897 ⁺	0.107	.836	.942 ⁺	.906 ⁺	.828 ⁺	.812	0.105	.807 ⁺	.376 ⁺	.964 ⁺	0.324	.778 ⁺	0.093	0.153	.524 ⁺	.402	0.030	0.194	.885 ⁺	.0001	.679 ⁺	.014	1.000	.040
	Si g _e	0.000	0.000	0.000	0.000	0.000	0.645	0.000	0.000	0.547	0.000	0.000	0.000	0.000	0.000	0.556	0.000	0.028	0.000	0.062	0.000	0.601	0.386	0.001	0.018	0.866	0.273	0.000	0.997	0.000	0.938		0.823
Ppmd p10min	r	.0094	.0094	.0054	.0092	0.043	.903 ⁺	.0091	.0077	.878 ⁺⁺	.0079	.0089	.0035	.0046	.0207	.925 ⁺	.0008	.474 ⁺⁺	.0021	.460 ⁺⁺	.0206	.914 ⁺⁺	.906 ⁺⁺	.261	.385 ⁺	.971 ⁺⁺	.203	.0187	.986 ⁺	.0024	.0126	.0040	1.000
	Si g _e	0.596	0.597	0.763	0.606	0.808	0.000	0.609	0.667	0.000	0.658	0.615	0.846	0.795	0.241	0.000	0.965	0.005	0.908	0.006	0.243	0.000	0.000	0.136	0.025	0.000	0.250	0.291	0.000	0.894	0.476	0.823	
	N	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34

Tabela 46 - Teste para igualdade de distribuições da base de dados total – Teste Mann-Whitney. Fonte: elaboração própria.

Variáveis	U e Mann-Whitney	Wilcoxon W	Z	p-value
Low Frequency	554	764	-4.01824	0.00005863
ApEn 0.10 50min	609	819	-3.70521	0.000211217
Very Low Frequency	611	8612	-3.69383	0.000220905
SampEn 0.10 50min	642	852	-3.51739	0.000435818
ApEn 0.10 u5min	657	867	-3.43201	0.000599118
SampEn 0.20 50min	678	888	-3.31249	0.000924692
SampEn 0.10 u5min	679	889	-3.3068	0.000943685
ApEn 0.15 50min	680	890	-3.30111	0.000963039
ApEn 0.20 50min	680	890	-3.30111	0.000963039
SampEn 0.15 50min	687	897	-3.26127	0.001109157
ApEn 0.15 p20min	700	910	-3.18728	0.001436195
ApEn 0.20 p10min	726	936	-3.0393	0.00237132
ApEn 0.10 u10min	740	950	-2.95961	0.00308025
ApEn 0.20 p5min	741	951	-2.95392	0.003137632
ApEn 0.15 p5min	749	959	-2.90839	0.003632954
SampEn 0.15 p20min	751	961	-2.89701	0.00376742
SampEn 0.10 u10min	755	965	-2.87424	0.004050008
SampEn 0.20 p10min	764	974	-2.82302	0.004757418
SampEn 0.10 u20min	765	975	-2.81732	0.004842556

ApEn 0.10 u20min	768	978	-2.80025	0.005106306
SampEn 0.15 u20min	769	979	-2.79456	0.005197067
ApEn 0.15 u20min	773	983	-2.77179	0.005574864
SampEn 0.15 u5min	773	983	-2.77179	0.005574864
SampEn 0.15 p5min	776	86	-2.75472	0.005874286
SampEn 0.10 p10min	780	990	-2.73195	0.006296048
SampEn 0.20 u5min	780	990	-2.73195	0.006296048
ApEn 0.20 p20min	782	992	-2.72057	0.006516987
ApEn 0.15 u5min	783	993	-2.71488	0.006630051
ApEn 0.10 p10min	784	994	-2.70918	0.006744875
ApEn 0.15 p10min	791	1001	-2.66934	0.007599959
ApEn 0.20 u5min	796	1006	-2.64089	0.008268952
SampEn 0.15 u10min	799	1009	-2.62381	0.008695192

Construção de um modelo utilizando a Análise Discriminante para a deteção de Acidemia Fetal

Tânia Raquel Martins Ferreira

FACULDADE DE CIÊNCIAS DA UNIVERSIDADE DO PORTO
FACULDADE DE MEDICINA DA UNIVERSIDADE DO PORTO

